DOCUMENT RESUME

ED 462 414 TM 033 676

AUTHOR	Hill, Clifford
TITLE	Linguistic and Cultural Diversity: A Growing Challenge to American Higher Education. CSE Technical Report.
INSTITUTION	California Univ., Los Angeles. Center for the Study of Evaluation.; National Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, CA.; College Entrance Examination Board, New York, NY.
SPONS AGENCY	Office of Educational Research and Improvement (ED), Washington, DC.
REPORT NO	CSE-TR-556
PUB DATE	2001-01-00
NOTE	68p.; Published in 2001 by the College Entrance Examination Board, here reproduced as No. 556 in the CSE Technical Report series. Sponsored by the National Task Force on Minority High Achievement, a 3-year initiative of the College Board.
CONTRACT	R305B60002
AVAILABLE FROM	UCLA/Center for the Study of Evaluation, 301 GSE&IS, box 951522, Los Angeles, CA 90095-1522 ($4.50). Tel: 310-206-1532. For full text: http://www.cse.ucla.edu/CRESST/Reports/TR556.pdf.
PUB TYPE	Reports - Descriptive (141)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	College Applicants; *College Entrance Examinations; College Preparation; *Cultural Differences; *Diversity (Student); *Educational Policy; Educational Research; *Higher Education; Minority Groups; Research Needs
IDENTIFIERS	*Scholastic Assessment Tests

ABSTRACT

This document originated in the Roundtable on Linguistic/Cultural Diversity and American Higher Education organized by the National Task Force on minority High Achievement held in San Juan, Puerto Rico, December 6-8, 1998. It represents one participant's view of important themes that emerged during the Roundtable. Three trends combine to present a growing challenge to U.S. higher education: (1) rapid growth in culturally diverse populations; (2) a substantial increase in the number of culturally diverse students seeking admission to higher education; and (3) a wide gap between these culturally diverse students and European American students in their performance on admission tests such as the Scholastic Assessment Test (SAT[R]). To respond appropriately to this challenge, the College Board will need to consider a broad range of policy options, some of which are described in this report. The first section focuses on preparatory programs for higher education. This section reviews two preparatory programs that the College Board has established, the Pacesetter[R] and the Advanced Placement Program (AP[R]) and recommends new directions for them to pursue to improve recruitment and instruction of culturally diverse students. The second section focuses on developing appropriate testing and assessment for culturally diverse students. It presents recommendations for improving the SAT and developing new approaches, which would not include versions of the SAT in other languages. It would be appropriate, however, to develop a modified English version of the SAT. The third section outlines a research

Reproductions supplied by EDRS are the best that can be made from the original document.

agenda that supplements the development of testing and assessment by improving databases of the College Board, conducting research on SAT performance for culturally diverse students, and focusing on research on culturally diverse students who perform at the highest levels on the SAT. This research should identify factors in the home and community, as well as in the school, that contribute to students' success. (Contains 3 figures, 9 tables, and 151 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

CRESST
National Center for Research on Evaluation, Standards, and Student Testing

ED 462 414

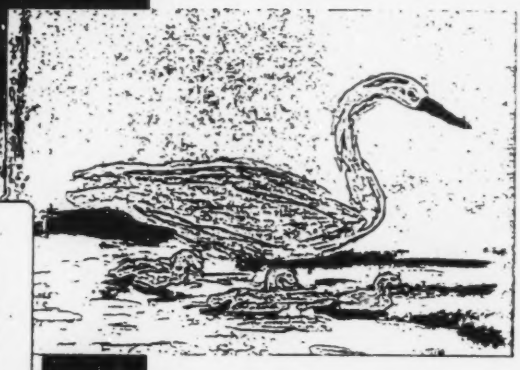
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
K. Hurst
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☐ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Linguistic and Cultural Diversity
A Growing Challenge to American Higher Education*

CSE Technical Report No. 556

Clifford Hill
Arthur I. Gates Professor of Language and Education
Teachers College, Columbia University
Visiting Scholar, The College Board



TM033676

UCLA Center for the Study of Evaluation
In Collaboration With

BEST COPY AVAILABLE

Linguistic and Cultural Diversity
A Growing Challenge to
American Higher Education*

CSE Technical Report No. 556

Clifford Hill
Arthur I. Gates Professor of Language and Education
Teachers College, Columbia University
Visiting Scholar, The College Board

January 2002

Center for the Study of Evaluation
National Center for Research on Evaluation,
Standards, and Student Testing
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1522
(310) 206-1532

* This report was published in 2001 by College Entrance Examination Board and is reproduced here as No. 556 in the CSE Technical Report series.

This report was published in 2001 by College Entrance Examination Board and is reproduced here as No. 556 in the CSE Technical Report series.

The report is in part a product of the National Center for Research on Evaluation, Standards, and Student Testing's Project 3.5, Coherence and Collaboration: Equity and Technical & Functional Quality, Robert L. Linn, CRESST/University of Colorado at Boulder, Project Director.

The work reported herein was supported in part by the Educational Research and Development Centers Program, PR/Award Number R305B60002, as administered by the Office of Educational Research and Improvement, U. S. Department of Education.

The findings and opinions expressed in this report do not reflect the positions or policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, or the U. S. Department of Education.

The College Board

Linguistic and Cultural Diversity

A Growing Challenge to American Higher Education

A Report Sponsored by the National Task Force on Minority High Achievement

Clifford Hill
Arthur I. Gates Professor of Language and Education
Teachers College, Columbia University
Visiting Scholar, The College Board

The College Board: Expanding College Opportunity

The College Board is a national nonprofit membership association dedicated to preparing, inspiring, and connecting students to college and opportunity. Founded in 1900, the association is composed of more than 3,900 schools, colleges, universities, and other educational organizations. Each year, the College Board serves over three million students and their parents, 22,000 high schools, and 3,500 colleges through major programs and services in college admission, guidance, assessment, financial aid, enrollment, and teaching and learning. Among its best-known programs are the SAT®, the PSAT/NMSQT™, the Advanced Placement Program® (AP®), and Pacesetter®. The College Board is committed to the principles of equity and excellence, and that commitment is embodied in all of its programs, services, activities, and concerns.

For further information, contact www.collegeboard.com.

Copyright © 2001 by College Entrance Examination Board. All rights reserved. College Board, Advanced Placement Program, AP, Pacesetter, Prueba de Aptitud Académica, SAT, and the acorn logo are registered trademarks of the College Entrance Examination Board. PAA is a trademark owned by the College Entrance Examination Board. PSAT/NMSQT is a joint trademark owned by the College Entrance Examination Board and the National Merit Scholarship Corporation. Other products and services may be trademarks of their respective owners. Visit College Board on the Web: www.collegeboard.com.

Contents

Acknowledgments v

Overview 1

Introduction 4

Preparation of Culturally Diverse Students for Higher Education . . . 7

Pacesetter® 7

Current Status 7

New Directions 10

Advanced Placement Program® 11

Current Status 11

New Directions 12

Appropriate Assessment for Culturally Diverse Students 15

Short-term Improvement of the SAT® 15

A Broader Base of Test Development 16

Selective Review of Research 16

Policy Recommendations 19

Test Accommodations 20

Selective Review of Research 20

Policy Recommendations 24

Long-term Development of New Approaches 26

Use of New Technologies in Assessment 26

Assessment of Potential for Learning 29

Assessment of Sociolinguistic Repertoires 32

A Further Research Agenda for Culturally Diverse Students 35
More Effective Use of Extant Data Bases 35
New Research Projects 36
Predictive Validity of the SAT 37
Coaching for the SAT 38
Case Studies of Successful Students: Family, Community, and School 43

Conclusions 45

References 45

Acknowledgments

This document has its origins in the Roundtable on Linguistic/Cultural Diversity and American Higher Education, which was organized by Edmund Gordon, who has chaired the National Task Force on Minority High Achievement at the College Board. He is the Richard March Hoe Professor Emeritus of Psychology and Education at Teachers College, Columbia University, where he also served as Dean.

I would like to acknowledge the mentoring role that Ed has played not only in the preparation of this document but also throughout my career at Teachers College. From the time I first joined the faculty, he has provided crucial support. As the Director of the Institute on Urban and Minority Education, he arranged for me to participate in a research seminar on cultural diversity and assessment that was jointly based at Teachers College and the Educational Testing Service. That seminar opened up a domain of research that doctoral students and I have pursued throughout the years. On the international front, this research was brought together in *From Testing to Assessment: English as an International Language* (Hill & Parry, 1994), and on the domestic front it led to *Children and Reading Tests* (Hill & Larsen, 2000). Ed, as a scholar and activist, has provided support in helping us move this research forward.

I would also like to express my appreciation to the other participants in the Roundtable, which was held in San Juan, Puerto Rico, December 6–8, 1998. The following individuals provided stimulating ideas, and a number of them made insightful suggestions on earlier versions of this document.

Participants from secondary and higher education*

John Baugh, Stanford University
Carol Bonilla Bowman, Ramapo College
Marta Cruz de Jésus-Janzen, Metropolitan State College of Denver
Eugene Garcia, University of California, Berkeley
Virginia Gonzalez, Texas A & M University
Gerunda Hughes, Howard University
Ruben Carriedo, San Diego City Schools
Evangeline Stefanakis, Harvard University

Participants from the College Board/Educational Testing Service*

Howard Everson, Vice President for Academic Initiatives
Manuel Maldonado Rivera, Executive Director, Puerto Rico Office
Scott Miller, Director, National Task Force on Minority High Achievement
Julia To Dutka, Executive Director, Test of English as a Foreign Language

Thanks go to Celeste Trinidad, who has served as an administrative and research assistant for the National Task Force on Minority High Achievement. She provided excellent support not only at the Roundtable but also during the months that followed.

This document is not a summary of what went on at the Roundtable, but rather my own view of important themes that emerged during our discussion. What I have chosen to put forward would not necessarily be endorsed by all the participants, since our Roundtable—like any productive human exchange—moved forward through the expression of different ideas and opinions. Many thanks to my fellow participants for the varied points of view they brought forth with vigor and conviction.

I am grateful to the College Board for the invitation to be a Visiting Scholar, which has provided an opportunity for me to reflect on the Roundtable discussion and to explore certain themes in greater depth. The personal warmth of Gaston Caperton, the president of the College Board, has made me feel at home, and many individuals have been helpful in providing information for this document. I would like to point out that the College Board, much like a university, has created an atmosphere that encourages independent thinking. Hence what I express in this document should not be viewed as reflecting official policies, but rather recommendations for the College Board to consider as it seeks to provide strategic support for the increasing number of culturally diverse students who are seeking admission to higher education.

I would like to thank the following individuals for providing suggestions on earlier versions of the manuscript: Jamal Abedi, Natasha Causton, Wade Curry, Pascal Forgione, Charlotte Gill, Arnold Goldstein, Steve Green, Akbar Herndon, Marie Keem, Lisa Larson,

* The institutional affiliations listed for the participants are the ones they held in December 1998, at the time of the Roundtable.

Maria Liu, John Olson, Martin Orland, Galina Savukova, and Maud Seghers. Particular thanks go, as always, to Eric Larsen, longtime friend and colleague at Teachers College, for his wise counsel and clear thinking about the issues under consideration.

Finally, I would like to express appreciation to the College Board and the Center for Research on Evaluation, Standards, and Student Testing (CRESST) for supporting not only the Roundtable but also the publication of this document. In particular, Howard Everson, Diane Foster, and Ron Dietel have been helpful in reviewing both the print and the online versions.

Overview

Three trends combine to present a growing challenge to American higher education:

1) a rapid growth in culturally diverse populations within American society
2) a substantial increase in the number of culturally diverse students seeking admission to higher education
3) a wide gap between culturally diverse students and European American students in their performance on admission tests such as the Scholastic Assessment Test (SAT®)

In order to respond appropriately to this challenge, the College Board will need to consider a broad range of policy options, some of which are presented in the three sections of this document. The first section focuses on preparatory programs for higher education. Many culturally diverse students, given the historical legacy of poor schooling, have not been in a position to perform well either on the SAT or in rigorous programs of higher education. Hence the College Board will need to maintain strong leadership in developing national programs that provide academic preparation for these students. The initial section reviews two preparatory programs that the College Board has set up—the Pacesetter® program and the Advanced Placement Program® (AP®)—and recommends new directions for them to pursue in order (1) to strengthen the recruitment of promising culturally diverse students, and (2) to improve the curriculum and instruction for these students.

The second section focuses on developing appropriate testing and assessment for culturally diverse students. It presents recommendations for both (1) improving the SAT, and (2) developing new approaches to testing and assessment. Given certain inherent limitations in current testing practices, improving the SAT can only be a short-term solution. Features such as speededness, arbitrary language complexity, and the multiple-choice format conspire to work against successful performance of students who use languages and dialects other than standard English in their own families and communities.

Given the research literature on test accommodations for culturally diverse students, the College Board should resist any pressure to develop versions of the SAT in other languages (or, for that matter, bilingual versions of the test). Such a policy is theoretically indefensible and fiscally imprudent. The College Board should, however, explore the possibility of using what accommodation researchers describe as "modified English" (i.e., language that avoids unnecessary complexity as a means of discriminating among students), since research indicates that all students—but especially culturally diverse students—perform better when linguistic complexity is minimized in text directions and questions.

As for long-term development of new approaches to assessment, the College Board will

need to focus on using computer technologies to develop tasks that are responsive to the new ways of processing information fundamental to a modern society. Students using these new technologies would no longer focus on extracting bits and pieces of information from written language alone. Rather they would use the multimedia capacities of the computer—integrating text with talk and visual imagery—in order to engage in tasks that assess a wider range of abilities. Just as research has shown that all students—but especially culturally diverse students—perform better when arbitrary language complexity is avoided, so an appropriate use of the new technologies could lead to a similar result, assuming that culturally diverse students are provided sufficient access to the technologies (assuring such access remains a major challenge to our educational system). Given this potential for diminishing the gap between culturally diverse students and European American students, the College Board will need to make a strong commitment to using computer technologies to move assessment in new directions.

At the same time, the College Board might investigate new ways of assessing culturally diverse students. Since these students often receive inferior schooling, assessment based only on academic achievement may underestimate their learning potential. If valid and reliable approaches can be developed to assess such potential, the educational system will be better positioned to nurture students who are capable of strong performance both in school and the larger society. Another possibility is to identify and assess particular strengths that these students possess. The sociolinguistic repertoires that they develop in negotiating their various worlds of experience represent resources, which, if properly developed, could enable these students to play vital roles in an increasingly global society.

The third section lays out a research agenda that supplements the development of testing and assessment. It begins with recommendations for improving extant databases at the College Board. These databases could be redesigned so that information is more readily available about culturally diverse students with respect to such features as their families' country of origin, educational background, and socioeconomic status. In implementing a new design, the College Board might consider a strategic collaboration with the National Center for Educational Statistics (NCES).

The College Board should continue to conduct research on two controversial issues with respect to culturally diverse students: (1) the degree to which the SAT accurately predicts their future academic performance, and (2) the degree to which their SAT performance can be improved through coaching. Previous research on these issues has been limited in various ways. With respect to predictive validity, researchers have not been sufficiently attentive to the degree to which the academic performance of these students varies with the quality of support that an institution of higher education provides (research has demonstrated that high-quality support is especially crucial at institutions that have historically served mainly European American students).

With respect to coachability, researchers have not sufficiently focused on culturally diverse students who cannot afford out-of-school coaching. Such students—especially those who are academically motivated—may be in a position to benefit significantly more from such coaching than students whose backgrounds ensure that they have already internalized a good deal of the knowledge and skills that it provides. The College Board should explore linking research not only to its delivery of test preparation materials through its Web site, but also to the emerging programs that target these students in various states (e.g., the College Preparation Program in California).

In addition, the College Board might consider conducting research on culturally diverse students who perform at the highest academic levels in its college preparatory programs. This research should be designed to identify various factors in the home and community as well as in the school that contribute to these students' success. Such research would be useful not only in improving policy and practice but also in providing effective role models for aspiring students and their families. It could also serve as an effective tool in recruiting such students into these programs.

Working with the National Task Force on Minority High Achievement, the College Board is in a position to develop strategic alliances with organizations such as the Center for Research on Evaluation, Standards, and Student Testing (CRESST). The issues that the College Board faces are formidable, but it can address them successfully if it exercises national leadership in mobilizing relevant organizations to work together. Through strong leadership, the College Board can help an increasing number of culturally diverse students not only to gain admission to institutions of higher education but also, once admitted, to perform successfully.

The ultimate beneficiary of increased participation of these students in higher education is the society as a whole. All students in American higher education can benefit from greater contact with cultural diversity, since communicating across cultural boundaries can help them come to a deeper understanding of what they are learning. Moreover, as culturally diverse students develop their language knowledge and skills in relation to specialized fields, they can play a strategic role in promoting international exchange in educational, cultural, and commercial spheres.

Introduction

In 2000, according to the United States Census Bureau, 37 percent of the population 17 years old and under were African American, Asian American, Hispanic/Latino American, or Native American. Throughout this document, individuals belonging to these four groups will be collectively referred to as *culturally diverse*.[1] Table 1 shows the distribution of this younger population across the four groups in 2000 as well as projections for the years 2010 and 2030.

TABLE 1

Proportion of Population 17 and Under Belonging to Four Culturally Diverse Groups

Group	2000	2010	2030
African American	15%	16%	16%
Asian American	4%	6%	7%
Hispanic/Latino American	17%	19%	25%
Native American	1%	1%	1%
Total	37%	42%	49%

As can be seen, the Hispanic/Latino American group is now larger than the African American group. This gap is projected to increase significantly because of greater immigration and a higher birth rate. By the year 2030, Hispanic/Latino Americans are projected to constitute a quarter of this younger population in the country as a whole,[2] and the four groups taken together will constitute nearly half of this population.

This projected increase in the culturally diverse population will be reflected in higher education enrollment. In a study conducted by the Educational Testing Service (Carnevale & Fry, 2000), the number of these students is projected to increase by about 2,000,000 during the next 15 years. Table 2 contrasts the proportion of the four culturally diverse groups of the total enrollment in 1995 with the projected proportion in 2015.

[1] The term *culturally diverse*, strictly speaking, refers to all groups, including European Americans; here, however, I will use it in a more limited sense to refer to those segments of the population whose primary identity is not European. I have simplified the descriptive terms used by the Census Bureau to characterize the four groups (for example, in the 2000 census the following subcategories were available for those who identified themselves as Hispanic: Mexican, Puerto Rican, Cuban, and other). Within American society, cultural identity for many individuals is best described with reference to more than one category, and for the first time the 2000 census made this possible. Only about 6,800,000 individuals (less than 3 percent of the total population) used this option; among those who did, the great majority (93 percent) used only two categories, and nearly four-fifths of this group chose White plus a second category. Until more individuals use this option, we are forced to rely on the broad, but potentially stereotypical, categories that permeate everyday discourse (see Bhopal & Donaldson, 1998, for a discussion of problems with these categories).

TABLE 2

Proportion of Higher Education Enrollment Belonging to Four Culturally Diverse Groups

Group	1995	2015
African American	13%	13%
Asian American	5%	8%
Hispanic/Latino American	11%	15%
Native American	1%	1%
Total	30%	37%

Although an increase of about 400,000 African American students is projected for 2015, they will remain at 13 percent of the total population and thus be surpassed by Hispanic/Latino American students who, along with Asian American students, are projected to increase at a significantly greater rate.

As the College Board plans for the future, one of its major challenges lies in developing a strategic approach to culturally diverse students. Table 3 shows the comparative performance of culturally diverse students in relation to European American students for the years 1990 and 2000 on the SAT.[3]

TABLE 3

Comparative Performance on the SAT (1990/2000)

Group	Verbal			Math		
	1990	2000	Gain	1990	2000	Gain
African American	428	434	6	419	426	7
Asian American	483	499	16	546	565	19
Hispanic/Latino American	459	461	2	464	467	3
Native American	466	482	16	468	481	13
European American	519	528	9	515	530	15
All	500	505	5	501	514	13

Although all groups have shown gains during this ten-year period, those registered by African American and Hispanic/Latino American students lag well behind those registered by the other groups.

---

[2] The Asian American group, much smaller than the Hispanic/Latino American group, grew at an even more rapid pace during the past decade: since 1990 the former group has increased 43 percent whereas the latter has increased 38.8 percent ("Hispanic and Asian Populations Expand," 2000, p. A16).

[3] For the sake of simplicity, I am using the acronym SAT to represent both the Scholastic Aptitude Test that was given in 1990 and the Scholastic Assessment Test I that was given in 2000. In comparing the performance of students at different points in time, these two tests are often equated, since they are each divided into two major components: verbal and math. By way of contrast, the Scholastic Assessment Test II focuses on major content areas such as history or biology. I would like to further note that the separate categories that the College Board includes for Mexican American and Puerto Rican students are not provided in Table 3. The names used for the various groups also differ slightly from those used by the College Board (e.g., *European American* is used rather than *White*).

As public attention becomes more sharply focused on the lower performance of these students, the SAT is coming under increasing scrutiny. At a conference sponsored by the Board of Testing and Assessment of the National Academy of Sciences in 1999, an entire session was organized around strategies that admission officers can use in dealing with various lines of attack against their continued reliance on admission tests such as the SAT. Two of the major lines are (1) its rather modest predictive validity of college performance (Bowen & Bok, 1998; Heubert & Hauser, 1999), and (2) the increased use of coaching by those who can afford it. In an article in *New York* magazine, Niles Lanning estimates that "nearly 100 per cent of Manhattan private-school juniors and seniors are paying for some kind of extracurricular SAT preparation" (1999, p. 32).

The proposal by Richard Atkinson, President of the University of California, that the SAT no longer be required is likely to lead to even more widespread debate around the role of standardized testing in higher education admissions. As reported in the *New York Times*, the proposal was stimulated by his observation of test preparation in a high school. As he puts it, "the time involved was not aimed at developing the students' reading and writing abilities but rather their test-taking skills. I concluded what many others have concluded—that America's overemphasis on the SAT is compromising our educational system" (Schemo, 2001, p. A13).

Clearly it is a strategic moment for the College Board to give serious consideration to developing new assessment policies with respect to culturally diverse students, and the College Board and CRESST are to be commended for sponsoring a Roundtable on Linguistic/Cultural Diversity and American Higher Education. In this document, which has been an outgrowth of the Roundtable, I outline a range of policies for the College Board to consider. In organizing this document, I have been guided by two points that were emphasized by Edmund Gordon at the Roundtable. First, no matter how important testing and assessment issues are, they should not mask the deeper issue of whether the educational system provides culturally diverse students with the academic preparation they need for effective participation in higher education: unless the system maintains high delivery standards, it is unrealistic to insist on high performance standards (see Gordon & Bonilla-Bowman, 1994, and Stewart & Everson, 1994, for further discussion). Second, it is crucial that academically talented students be identified and provided a strong challenge early on in their education. Such a policy is important in developing individuals who will eventually exercise intellectual leadership not only within the academy but also within other professions. From a long-term perspective, such leadership is vital in facilitating effective integration of culturally diverse populations into American society.

Given these two guidelines, I first address how the College Board can exercise national leadership in recruiting culturally diverse students who are academically talented for college preparatory programs. I review two major programs developed by the College Board and suggest ways in which they might be strengthened to better serve these students.

I then discuss various areas that the College Board might explore as it seeks to develop appropriate assessment for culturally diverse students. I begin this section by focusing on short-term improvement of the SAT, with particular attention to accommodation issues. I then shift to long-term development of new approaches to assessing these students, initially exploring how the new technologies might be used to transform the intellectual framework of assessment and open up new ways of evaluating a wider range of abilities. I end this section by discussing ways in which these students' potential for learning as well as their sociolinguistic repertoires might be assessed.

Finally, I lay out a research agenda that extends beyond the development of appropriate testing and assessment. I begin this section by exploring how the extant databases at the College Board might be used to provide more refined information about culturally diverse students. I also suggest ways in which more informative research can be conducted on the issues of predictive validity and coaching with respect to these students. I end this section by suggesting case studies of academically successful students that identify factors in the home environment as well as in the school environment that contribute to their high level of performance.

Preparation of Culturally Diverse Students for Higher Education

This section explores various ways in which the College Board can strengthen its precollegiate programs so that they can more effectively serve culturally diverse students who show academic promise. I discuss first the Pacesetter program and then the Advanced Placement Program.

Pacesetter®

Current Status

The Pacesetter program was introduced in 1993 with a senior-level course in mathematics, and in the following year it established courses in English and Spanish. These courses were designed to reflect

1) high academic standards
2) innovative curricula
3) balanced testing and assessment

In developing these courses, the Pacesetter program drew on the national standards sponsored by the relevant professional organizations. The Spanish course, for example, was based on the *Standards for Foreign Language Learning* published by the American Council on the Teaching of Foreign Languages (ACTFL). ACTFL standards were used to

develop three basic goals for Pacesetter students, which can be conveniently placed under the categories of communication, culture, and connection:

1) Students should learn to communicate in Spanish.
2) Students should use Spanish to gain knowledge of Hispanic/Latino cultures.
3) Students should use Spanish to connect with disciplines and thus acquire academic knowledge through the language.

The innovative curriculum is exemplified in the first unit, *¿Nos Conocemos?* [Do We Know Each Other?], which focuses on immigration of Spanish-speaking people to the United States. Students are presented four kinds of material within this unit:

1) a recorded interview with a Mexican immigrant
2) oral and written versions of a poem by an immigrant
3) a Spanish-language newspaper published in New York City
4) an article on ethnicity and immigration presenting graphs and charts of census data

Students engage with this material through various individual and small-group activities that culminate in a project: a biographical (or autobiographical) presentation in which they are encouraged to integrate different media such as print, photography, and videotape.

The testing and assessment component of the course is designed to provide both formative and summative evaluation. Throughout the course students maintain a portfolio of learning logs and related journal activities that the teacher can use for diagnostic purposes. In addition, they engage in individual and small-group activities in which they carry out various kinds of self- and peer-assessment. At the end of the course, students are given a nationally administered exam that calls for constructed responses.* The exam is presented in four sections—speaking, listening, reading, and writing—and is thematically related to the curriculum. Here are sample activities from the 1998 exam:

1) Speaking: Students are presented graphs showing the population growth of Spanish-speaking people in the United States. They orally describe what the graphs represent.
2) Listening: Students listen to an excerpt of a university lecture about the settlement of Spanish-speaking people in the United States. They then place various events under three columns headed *Mexicanos, Puertorriqueños*, and *Cubanos*.
3) Reading: Students read a passage about Spanish newscasting and then answer questions about distinctive properties of such newscasting (there is also an applied activity in which students imagine they are the producer of a 30-minute newscast in Spanish; they are asked to provide the content and duration of various components within this newscast).

* One commendable feature of the Pacesetter examination system is its avoidance of multiple-choice questions: they are used only on a limited portion of the mathematics exam that deals with calculations.

4) Writing: Students are asked to carry out two writing activities:
   (a) a letter to a commercial company that has erroneously treated the Hispanic/Latino community as homogenous (they are to explain not only the error but also its implications for the company)
   (b) a description of a project (e.g., a video or a festival) designed to show the diversity of the Hispanic/Latino community in the United States

The final exam has been presented in some detail, since it embodies a number of attractive features (e.g., students construct responses to the kinds of tasks that they have worked with in the classroom). One potential problem—and it is often evidenced in testing built on the new standards in American education—is that it requires students to carry out complex activities in a limited amount of time: for example, students are provided only 40 minutes in which to do the two writing activities. Here are the directions they are to follow while engaging in these activities.

> **Instrucciones:** Contesta las siguientes preguntas de la forma más completa posible. La evaluación tomará en cuenta el uso apropriado del vocabulario, la expresión creativa, la organización y su relevancia temática.
>
> [**Directions:** Answer the following questions as fully as possible. Your work will be evaluated for appropriateness of vocabulary, creativity, organization, and thematic relevance.]

It is not at all clear just how students can manage within 40 minutes to write "as fully as possible" while being creative, well-organized, and thematically relevant on two separate topics. When complex writing activities are introduced, students need to be allotted sufficient time, or perhaps the time restriction should be removed altogether. Students can carry out writing projects apart from a timed testing situation and include them in a portfolio along with brief samples of writing produced in a testing situation (the writing produced on a test provides a useful check to insure that the other writing represents the student's own work).

The three components of the courses—high academic standards, innovative curricula, balanced testing and assessment—have been combined with intensive teacher-preparation courses to insure the successful launching of the Pacesetter program. One measure of success has been its rapid growth. The program now enrolls over 55,000 students: 28,930 in English, 17,129 in mathematics, and 9,148 in Spanish. Moreover, the number of schools participating in the program has grown rapidly: the mathematics course is now in 320 schools, the English in 272 schools, and the Spanish in 142 schools.

Another measure of success can be found in an evaluation conducted by the Human Resources Research Organization (HRRO). Using a quasi-experimental design, HRRO compared students in the Pacesetter English course with those in a regular English course

at various sites around the nation. The study showed that the Pacesetter students performed better not only on the Pacesetter final exam but also on the NAEP reading test and the AP writing test.

New Directions

At this point, the number of culturally diverse students in the Pacesetter program cannot be determined, since it does not yet collect this background information. According to its director Steve Green, there are plans to gather such data in the near future. These plans should be acted upon as soon as possible. It is important that the program collect information not only on race/ethnicity, but also on country of origin, home language(s), family income, parental education, and amount of time spent in this country. Such information will be crucial in mounting more intensive recruitment of promising students, especially those who cannot afford academic preparatory programs.

One major way of increasing enrollment of culturally diverse students is to show that they perform successfully within the Pacesetter program. If the HRRO study, for example, had documented a strong performance by these students in the English course, this information could have been used to encourage schools with substantial culturally diverse populations to introduce the course. In a later section of this report, I suggest that the College Board might consider conducting case studies of successful students in its preparatory programs that focus on important factors in the family, community, and school that contribute to success. These case studies could be particularly effective in increasing enrollment of culturally diverse students in the Pacesetter program.

The Pacesetter program has two other kinds of plans that are potentially beneficial to culturally diverse students. First, the program is considering the development of new courses that would effectively serve these students. One course under consideration is World History. Such a course would provide these students an opportunity to explore their own cultural heritage within a larger intellectual framework (much as the multicultural curriculum of the English course is designed to encourage such exploration). Second, the program is introducing earlier levels of existing courses, which will be particularly beneficial for culturally diverse students, since Pacesetter courses are currently introduced relatively late in their academic career. As researchers (Collier, 1989; Cummins, 1981) have observed, students who have immigrated to this country minimally need about 6 to 7 years to develop their academic knowledge and skills in English (for those who have not developed an academic foundation in their first language, this period can be even longer). The College Board should give serious consideration to developing a more comprehensive Pacesetter program to be introduced even as early as the beginning of middle school.

Advanced Placement Program®

Current Status

The Advanced Placement Program (AP) was developed on the premise that college-level courses can be successfully taught to academically talented students at the secondary level. Since its introduction in 1955, it has grown steadily, and nearly 60 percent of the nation's 22,000 high schools now offer AP courses. In 1999–2000, more than 625,000 students took 1,242,324 AP Exams in 32 different subjects ranging from calculus to studio art. Table 4 shows the number of these exams and the proportions of the total that were taken by various groups of students. Asian American students are, by far, the heaviest consumers of AP courses.[5] As for the other groups, Hispanic/Latino students are fairly well represented, since many of them take the Spanish language exam (to be discussed below). It is African Americans and Native Americans who are the most underrepresented in AP courses.

TABLE 4

Number of AP® Exams Taken by Various Groups of Culturally Diverse Students

Group	Number	Percentage
African American	53,136	4.3%
Native American	5,449	0.4%
Asian American	167,490	13.8%
Mexican American	59,720	4.9%
Puerto Rican	7,546	0.6%
Other Hispanic/Latino	45,293	3.7%

Table 5 shows the mean grade for culturally diverse students and European American students in all AP subjects (grades are on a scale of 1-5, with 5 the highest grade). The overall performance of the three Hispanic/Latino groups was bolstered by a high rate of participation in the Spanish language course. Table 6 shows the number of students in these three groups who took the Spanish language exam and their mean grade (these mean grades were well above the overall mean grade of 3.66). Eugene Garcia (1994) points out that this strong performance of Hispanic/Latino students in the Spanish language course has not only motivated many of them to enroll in institutions of higher education but also provided them with a sense of self-confidence that they can perform successfully at this level.[6]

[5] In calculating the proportions of the total exams, I have excluded the 29,271 exams taken by students who did not identify their ethnocultural background.

[6] Nearly half of the Spanish language exams are taken by Hispanic/Latino students.

TABLE 5

Mean Grade in AP Courses for Various Groups of Students

Group	Mean Grade
African American	2.17
Native American	2.62
Asian American	3.06
Mexican American	2.77
Puerto Rican	2.83
Other Hispanic/Latino	3.09
European American	3.06
All	3.01

TABLE 6

Number of Spanish Language Exams and Mean Grade for Hispanic/Latino Groups

Group	Exams	Mean Grade
Mexican American	16,435 (26.7%)	4.33
Puerto Rican	1,331 (1.8%)	4.12
Other Hispanic/Latino	11,748 (18.5%)	4.52

New Directions

The College Board is considering the possibility of developing Spanish language exams for their AP courses in biology and calculus (and even for world history if such a course is developed). If the College Board does move in this direction, it would provide additional courses where Hispanic/Latino students can strengthen their academic performance by using their first language. I would, however, like to issue a word of caution: as the research of Abedi, Lord, and Hofstetter (1998) indicates, students will not necessarily be in a position to perform better on a calculus exam given in Spanish, unless the language has also been used as the primary vehicle of instruction. Hence if the College Board does move forward in this area, it will need to insure that both teaching and testing will be conducted in Spanish.

As for the performance of African American students in AP courses, there is no particular course—like the Spanish language course for the Hispanic/Latino students—in which they can register a strong performance by virtue of their linguistic heritage. It is interesting to observe, however, that their strongest performance was generally in language and literature courses with the exception of English language and English literature. Table 7 shows the mean grade of African American students in all language and literature courses. The relatively poor showing in the English courses may well be related to what linguists (e.g., Baugh, 1983; Labov, 1972; Smitherman, 1975) have often pointed out: the

English used by many African American students in their homes and communities provides considerable interference in developing effective control over the English required for successful school performance (unlike, say, Spanish-speaking or Chinese-speaking students, they do not have the resource of a distinctly separate language to use in their homes and communities). Such dialect interference is particularly unfortunate, since these students, as Table 7 indicates, do show their greatest strength in language and literature courses that do not involve English. In both English language and English literature courses, their mean grade (2.12) is below their overall mean grade (2.17), whereas in all the other language and literature courses their mean grade is well above their overall mean grade. For example, their mean grades in French language and German language were 2.65 and 2.87 respectively, which were quite close to the mean grades for all students taking these courses (2.75 and 2.98 respectively).

TABLE 7

Mean Grade for African Americans in Language and Literature Courses

Exam	Mean Grade	
English	language	literature
	2.12	2.12
French	language	literature
	2.65	2.77
German	language	
	2.87	
Latin	Vergil	literature
	2.38	2.53
Spanish	language	literature
	2.39	2.92

The complex sociolinguistic experience of these students may well develop their capacity for language learning, but this experience is not effectively drawn on in the school setting where the language to be learned is in direct conflict with the social dialect they speak in their homes and communities. As Ogbu (1994) has pointed out, this conflict involves deeper issues of ethnocultural identity that have to do with the historical experience of African Americans within the larger society. These issues are complex and warrant a more extended discussion than can be provided here. They are, however, of sufficient importance that the College Board might consider bringing together a group of scholars to explore them within the context of testing and assessment.

Given the relatively strong showing of African American students in foreign language and literature courses, the College Board might also consider exploring ways of encouraging them to participate in these courses. Successful performance in such courses can help build a sense of academic achievement among African American students (much

as Latino/Hispanic students build up self-esteem by successful participation in Spanish courses). The AP courses in literature should incorporate, wherever possible, works by African, Caribbean, and African American writers (e.g., Francophone African literature has been effectively incorporated in the French literature course).

Finally, the AP program should consider introducing new foreign languages. At the present time, it is exploring the possibility of introducing Chinese, Japanese, Italian, and Russian. It might also consider the possibility of introducing other Asian languages—for example, Korean and Vietnamese—that are widely spoken in immigrant communities. Introducing Asian languages would be justified in light of the large number of Asian American students enrolled in AP courses.

I would like briefly to address the nature of the exams within the AP program. These exams consist of two parts: multiple-choice questions and a written essay. Since these courses are offered for college-level credit, the College Board might explore the possibility of developing a broader assessment model. Since college courses often include a research project conducted throughout the trimester/semester, it would increase the face validity of AP courses to include a research project that accounts for at least some portion of the course grade.

The College Board might consider consulting with the International Baccalaureate about the most efficient way of conducting large-scale evaluation of course projects. For a number of years now, the International Baccalaureate has been preparing teachers from a range of ethnocultural backgrounds to function as external assessors of such projects. These assessors are sensitized to language and culture issues since they are dealing with projects written in English by students from countries throughout the world. Assessors are generally advised, for example, not to penalize non-native speakers of English for use of nonstandard language if it does not interfere with either clarity or content. Central to the preparation of assessors is the use of exemplar projects to anchor the marking schemes (see Hill, 1996, for three criteria—*clarity, content, critical thinking*—used to evaluate exemplar projects for the innovative course *Theory of Knowledge*).

Within such a broad model of assessment, the written essay should be retained in the final exam, since it provides an important means of checking that individual students have, in fact, produced their own course projects (any major inconsistency between the course project and the written essay becomes readily apparent). The use of constructed responses within the final exam should also be explored (see, for example, the Pacesetter final exam), since culturally diverse students are disproportionately represented among students who perform poorly on multiple-choice tasks. As artificial intelligence programs become increasingly reliable in scoring constructed responses, the rationale for using multiple-choice tasks is much less compelling (see Bennett, 1999, Sebrechts, Bennett, & Rock, 1991; Wang & Hill, 1999, for discussion of the strategic role that artificial intelligence programs can play in scoring constructed responses).

Before leaving this discussion of precollegiate programs, I would like to commend the College Board for its work in preparing teachers to participate successfully in these programs. Historically, the College Board has played an important role in American education in facilitating cooperation between secondary and tertiary educators in developing strong preparatory programs. I also commend the College Board for its recent efforts to expand teacher education within its preparatory programs, for ultimately the strength of these programs will be determined by how well the teachers are prepared. As the College Board engages in such expansion, it needs to maintain a strong focus on preparing teachers to work effectively with culturally diverse students. To achieve this end, the College Board might develop collaborative relationships with graduate schools of education that have a rich history of preparing teachers to work with these students. Through such collaboration, the College Board would be in a position to exercise national leadership in this area (the recent establishment of the Visiting Scholars Program can be viewed as an effective means of initiating such collaboration).

Appropriate Assessment for Culturally Diverse Students

In developing appropriate testing and assessment for culturally diverse students, the College Board will need to distinguish between long-term and short-term goals. The short-term goal is to explore various ways in which the SAT can be altered, especially given the increasing public scrutiny of this test. The long-term goal is to explore developing viable alternatives. In developing these alternatives, the College Board will need to explore how the new technologies can be most effectively used in assessing culturally diverse students' knowledge and skills. It might also explore promising areas such as assessing their potential for learning as well as their sociolinguistic repertoires.

Short-Term Improvement of the SAT

For the immediate future, the SAT will continue to play a strategic role in the admission decisions of American colleges and universities. As Howard Everson pointed out at the Roundtable, admission personnel are generally not interested in using a broader assessment model (e.g., one that formally incorporates student portfolios). As was pointed out in a *New York Times* Magazine article on the SAT, "for college officials seeking to distinguish thousands of students from different backgrounds and high schools of varying quality, and different curriculums and grading policies, the SATs do provide a common currency, whatever their predictive power." As one of these officers at Dartmouth College put it, "When you're at sea, you'll take anything that looks like a landmark" (Schwartz, 1999, p. 51).

In order to improve the SAT on a short-term basis, I consider first how the base of test development might be broadened to insure greater equity for culturally diverse students and then whether any of the proposed test accommodations should be adopted.

A Broader Base of Test Development

Selective review of research

During the past two decades, researchers have used the methods of discourse analysis to examine not only how test material is constructed but also how representative test takers, including culturally diverse students, respond to it.[7] Hill and Larsen (2000) identify three factors that are crucial in explaining how test material discriminates among students:

1) speededness
2) language demands
3) multiple-choice format

These factors interact in complex ways to insure that certain students—and culturally diverse students are disproportionately represented among them—perform relatively less well when taking a test, even though they may be capable of performing successfully in higher education.

In order to provide a sense of how these factors interact, let us consider actual material from the SAT. In dealing with its three verbal sections, students have 75 minutes to answer 78 questions. During this time, they must answer three different kinds of questions:

1) sentence completion (19 questions)
2) analogy (19 questions)
3) critical reading (40 questions)

In order to answer the critical reading questions, students must read 5 passages that deal with different topics. Here, for example, are the topics of the passages used on the January 1997 SAT:

1) a linguist's discussion of changes in English language usage
2) a visual artist's memories of her first visit to Paris
3) an anthropologist's account of cultural influences in the United States
4) an account of the astronomer Jansky discovering radio waves
5) the astronomer Reder's explanation of why astronomers were initially uninterested in this discovery[8]

Given the limited amount of time that students have, they can experience a good deal of difficulty in shifting among diverse topics. In order to become engaged with a given topic, they must activate an appropriate set of schemata that allow them to understand

---

[7] For examples of this research, see Adames, 1987; Allen, 1988; Aronson & Farr, 1988; Bhasin, 1990; Chu, 1993; Coyle, 1992; Freedle, 1979; Freedle & Duran, 1987; G. Garcia, 1988; Gonzalez, 1996; Gonzalez, Bauerle, & Felix-Holt, 1996; Haney & Scott, 1987; Hill, 1977a, 1977b, 1992, 1995, 1999a, 1999b; Hill, Anderson, Watt, & Ray, 1989; Hill & Larsen, 1983, 1992, 2000; Hill & Parry, 1988, 1989, 1992, 1994; Hill & Pike, 1985; Ingulsrud, 1988; Johnston, 1984; Langer, 1987; Nix & Schwartz, 1979; Parry, 1986; Pearson & Valencia, 1987; Sims-West, 1996; Yuan, 1997.

[8] The last two passages were presented together and were followed by a set of 13 questions: 5 based on the first passage, 4 on the second, and 4 on both.

what they are reading. Once they have answered the questions on a passage, they must let go of one set of schemata and immediately activate another set that will allow them to become engaged with the new topic. In research on test taking (e.g., Parry, 1986), students often talk about their difficulty in moving rapidly from one textual world to another, especially when the worlds hold little interest for them and high stakes are attached to whether they can become sufficiently engaged.

Students' difficulty in moving in and out of textual worlds is increased by the language demands that these worlds present. To take a closer look at these demands, let us return to the January 1997 SAT. To begin with, the texts (the passages and accompanying questions) were quite long. The longest text contained 1,459 words (the passage, 838, and the multiple-choice questions, 621). Moreover, 9 of the 12 questions called for sentence completion, and many of these contained long stems (e.g., 30 words) and long options (e.g., 9–12 words). I will shortly discuss how such long questions are especially difficult to process.

In addition, the passages contained a good deal of vocabulary that is not frequently used (e.g., *disinterested, improprieties, antiquated, nihilism, mutualism, intellegentsia*).[9] Finally, the passages were filled with long sentences that contain multiple clauses. Consider, for example, the final two sentences of the passage about changes in English language usage:

> What is otherwise a natural appeal to a shared background is distressing, because we do not know who the speakers are, as we do in face-to-face conversation, and we cannot ask them for clarification. Just as attention to rules of written usage helps us to read intelligently, so an awareness of the abuse of 'you know' in public forums makes us better listeners.

The first of these sentences contains 36 words in 6 clauses; the second contains 29 words in 2 clauses. Within each sentence, the internal structure of the clauses can be quite complex. In sentence 1, the initial clause, which functions as the subject of another clause, contains 10 words (i.e., *What is otherwise a natural appeal to a shared background*). In sentence 2, which is comparative in structure, the *just as*-initiated clause contains a subject of 6 words in 3 phrases (i.e., *attention to rules of written usage*), and the *so*-initiated clause contains a subject of 11 words in 4 chunks (i.e., *an awareness | of the abuse | of 'you know' | in public forums*). Lengthy subjects are difficult to process—the reader keeps waiting for the verb to appear—especially when they occur within a long sentence that contains many clauses. Such demanding language can place a good deal of strain on the working memory of test takers (see Kyllonen & Christal, 1990; Snow & Lohman, 1989). Such strain can be particularly severe for culturally diverse students whose working memory for complex forms of English tends to be less developed (just as the working memory of native speakers of English is generally less developed for complex forms of a second language they have learned).

---

[9] Infrequently used vocabulary was also present in the sentence-completion and analogy questions on the test (e.g., *circumlocutions, gibberings, machinations, deleterious, inviolable, evanescent, sagacious, querulous*).

The strain on working memory becomes even greater when students turn to the multiple-choice questions. Consider, for example, question 18, one of 5 that follow the passage about changes in English usage:

It can be inferred from the passage that the author approves most of modern users of language who

a) believe that meanings of words are purely arbitrary

b) treat public conversation as if it were private

c) recognize the reasons for particular usages

d) consider the natural drift of language to be inescapable

e) relax the rules of written usage.

The incomplete sentence that forms the stem of this question is 18 words long. It contains two clauses (the first is in the passive voice) and the beginning of a third that must be completed (it also contains an awkward placement of the word *most* that can be confusing to readers).

Since the options are themselves complex, integrating them with the incomplete stem can be demanding. Here, for example, is the number of words and clauses in the five sentences to be formed from a single incomplete stem:

a) 25 words and 4 clauses

b) 25 words and 4 clauses

c) 23 words and 3 clauses

d) 26 words and 3 clauses

e) 23 words and 3 clauses

None of these sentences is actually written out, so students are forced to continuously work back and forth between the question stem and the five options. While attempting to hold these virtual sentences in memory, they may also be forced to return to the passage several times as they evaluate which is the correct answer.

This task of working back and forth between virtual sentences and the passage is complicated by certain features of the passage. The passage itself, as we have seen, contains awkwardly constructed language, but even worse, questionable ideas. A brief introduction to the passage presents it as having been written by a linguist. Yet many of its ideas would be unacceptable to virtually any linguist (e.g., listeners will operate more efficiently if they become aware that the frequent use of *you know* in public speech is inappropriate).

In this report, I do not have space to explore how problematical aspects of content interact with test performance, but such interaction has been explored in depth in a number of studies (e.g., G. Garcia, 1988; Gonzalez, 1996; Hill, 1992; Hill & Larsen, 2000; Hill & Parry, 1994; Yuan, 1997). I would like to note, in passing, that perhaps the most important skill in critical reading is using one's knowledge base to think carefully about what one reads. If test takers exercise this skill, it can work against successful test

performance: not only does it take up too much time but it risks transforming the passage information to which they must restrict themselves.

The demands of multiple-choice questions are further complicated by the burden on test makers not only to insure that the keys are unassailable but also that the distractors are genuinely attractive (a multiple-choice question will not make it onto a test unless a certain proportion of test takers is attracted to its distractors during piloting).[10] In a study of multiple-choice questions designed to test reading comprehension, Hill and Larsen (2000) examined various ways in which test makers meet the dual demands of defensible keys and attractive distractors. One common way is that they juxtapose a key that deals with relatively unimportant information (but is technically accurate) with a distractor that deals with relatively important information (but is based on an inference, which, though stimulated by the passage, expands this information beyond what the test-taking situation will allow).

Since reading is a constructive process in which real-world schemata are used to make sense out of what is on the page, certain kinds of expansion are normative in ordinary acts of reading. Indeed, one of the major dimensions of the reading process is integrating what is on the page with what is already in the head. If students are, however, to be successful test takers, they must learn to suppress the inferences that lead to such integration. In their research, Hill and Larsen found that culturally diverse students were disproportionately represented among those who selected distractors that lead to passage expansion. Their research led them to challenge the use of the multiple-choice format, especially since it introduces an arbitrary task that has little to do with the normative processes of reading comprehension.

Policy recommendations

In order to deal with the problems that researchers have identified, the College Board should set up, in consultation with CRESST, an independent committee of scholars to broaden the base of test development. The current reliance on psychometric procedures of review provide important elements of rigor, but they inadvertently privilege the responses of the students who are the most effective test takers. Such privileging comes from reliance on the technique of biserial correlation to determine whether prospective material actually makes it onto the test. In essence, this technique insures that only items on which the most successful test takers perform well—and conversely, the least successful test takers perform badly—end up on the test. Since test makers must insure that a certain proportion of items discriminate sharply among test takers, they must select some items for which only the most successful test takers are consistently able to select the key (see Lucas, 1998, for further discussion of the ways in which current methods of test development work against the interests of culturally diverse students).

[10] A key is the answer designated as correct by the test makers, whereas distractors are the answers designated as incorrect.

In order to deal with the limitations of psychometric methods, the review committee would use the methods of discourse analysis to investigate not only prospective test material, but also how representative test takers respond to it. The committee could work within the framework that the Board of Testing and Assessment of the National Academy of Sciences has recommended for developing national tests in reading and mathematics. Prominent within this framework is the use of think-aloud protocols and retrospective interviews to document the responses of representative test takers, including appropriate samples of culturally diverse students, to prospective test material.

Within a research framework initiated by Hill and Larsen (1983) that has been used in this country (e.g., Sims-West, 1996) and abroad (e.g., Yuan, 1997), think-aloud protocols and retrospective interviews constitute the initial stage of a three-stage model of research:

1) protocols and interviews are used to discover language problems at both macro- and microlevels
2) the test material is rewritten in order to deal with these problems
3) the rewritten material is administered to representative test takers (including culturally diverse students) to determine whether their performance improves

This model has shown that students—especially culturally diverse students—improve significantly in their test performance when language problems are minimized. As the review committee carries out its work, it will need to pay particular attention to the ways in which these problems interact with speededness and the multiple-choice format.

Test Accommodations

Selective review of research

There is an emerging research literature about how best to provide test accommodations for culturally diverse students whose English language skills are not sufficiently developed.[11] This research has been largely centered on efforts by the National Assessment of Educational Progress (NAEP) to develop a more inclusive approach to the monitoring of educational achievement. This research has often linked two groups that were traditionally excluded from such assessment: students with disabilities and culturally diverse students who speak a language other than English.

In the case of students with disabilities, test accommodations are dictated by the nature of the particular disability: for visually impaired students, tests are, for example, administered in braille. In the case of culturally diverse students, however, the test

[11] For examples of this research, see Abedi, Lord, & Hofstetter, 1998; Abedi, Lord, & Plummer, 1995; Anderson & Olson, 1996; Anderson, Jenkins, & Miller, 1996; August & Hakuta, 1997; Butler & Stevens, 1997; Cheung, Clements, & Miu, 1994; Fleischman & Hopstock, 1993; Hakuta & Valdes, 1994; Holmes, Hedlund, & Nickerson, 2000; Hopstock & Bucaro, 1993; Houser, 1995; Menken, 2000; Olson, 1995; Olson & Goldstein, 1996, 1997; Rivera & Stansfield, 1998; Rivera, Vincent, Hafner, & LaCelle-Peterson, 1997; Secada, 1994; Shepard, 1996; Stancavage, Godlewski, & Allen, 1996; Zehler, Hopstock, Fleischman, & Greniuk, 1994.

accommodation has been more difficult to specify. Hitherto three major kinds of accommodations have been tried out:

1) use of modified English
2) use of the first language
3) use of the first language and English (i.e., a bilingual test)

Before reviewing research on these kinds of accommodations, I should note that each can be applied either to the entire test or to only a certain portion of it. In the case of modified English, for example, the instructions rather than the entire test may be simplified (often they are presented in speech as well as in writing). In the case of bilingual testing, the use of the students' first language may be used for only certain vocabulary items rather than the entire text. In experimental administrations of NAEP science and mathematics tests, a separate glossary of key terms is often provided in the students' first language (hitherto Spanish because of the large number of Hispanic/Latino students). In the case of English language tests for these students, the NAEP policy has been not to provide such glossaries.[12]

In a study of NAEP math items with culturally diverse students in the 8th grade, Abedi, Lord, and Hofstetter (1998) administered the items in

1) unmodified English
2) modified English
3) Spanish

In the case of the modified English version, "care was taken to avoid altering special mathematics vocabulary and structures; only the non-technical 'ordinary' language was modified" (p. 3). This ordinary language was modified with respect to

1) vocabulary frequency/familiarity
2) the complexity of various structures (e.g., prepositional phrases, relative clauses, conditional clauses, and comparative structures)
3) the length of certain units (e.g., words, noun phrases, sentences, and the test items themselves)

Figure 1 presents a math task in the unmodified version and the modified version in order to illustrate the effects of the language modifications. As can be seen, the math problem is simple, but the language in the unmodified version is fairly complex. It is as if language—not mathematical—comprehension is being tested.

[12] In many countries, the glossing of unfamiliar vocabulary items is common in the testing of English. In the People's Republic of China, where authentic texts are used in testing English, any unfamiliar vocabulary item is immediately followed by a gloss in Chinese.

FIGURE 1 Modified and Unmodified Versions of a Math Question

Unmodified English

If ☐ represents the number of newspapers that Lee delivers each day, which of the following represents the total number of newspapers that Lee delivers in 5 days?

(A) 5 + ☐
(B) 5 x ☐
(C) ☐ + 5
(D) (☐ + ☐) x 5

Modified English

Lee delivers ☐ newspapers each day. How many newspapers does he deliver in 5 days?

(Abedi, Lord, & Hofstetter, 1998, p. 3.)

According to the authors, the following changes were made in order to achieve greater simplicity in the modified version:

1) long nominals were shortened
2) two relative clauses were removed and recast
3) the initial question phrase was changed from "which of the following represents" to "how many"
4) the number of clauses was reduced from 4 to 2
5) the average number of clauses per sentence was reduced from 4 to 1
6) the average sentence length was reduced from 26 to 6.5 words
7) the item length was reduced from 26 to 13 words

The authors report significant differences in culturally diverse students' performance on the three versions: students performed highest on the modified English version, lower on the unmodified English version, and lowest on the Spanish version. With respect to the modified and unmodified English versions, a control group of European American students also performed better on the modified English version than on the unmodified version. The improvement in their performance, however, did not match that of the culturally diverse students. The authors thus recommend that the language used in tests be carefully monitored to insure that it is not used as a tool to discriminate among students.

As for the students' lower performance on the Spanish version of the test, the authors suggest that this result is related to the fact that English is the dominant language of math instruction for most of the students. As evidence for this interpretation, they point out that the performance on the Spanish language version of the test was significantly higher for the small number of students who had been taught math in Spanish. They thus recommend that the language used in assessing these students should be the language in

which they have been instructed.[13]

As for the accommodation for culturally diverse students based on a bilingual test, NAEP conducted a field test in math with Spanish-speaking students in the 8th grade in 1995. According to Anderson, Jenkins, and Miller (1996), the students were instructed to answer either the English or Spanish version, but not both versions. In general, most students followed directions, with the vast majority of them (more than 80 percent) answering only the Spanish version. A statistical analysis of student performance indicated that for many items the measurement properties were not parallel. It was thus difficult to build an integrated psychometric scale for the two versions of the test. The authors hypothesize that the lack of psychometric parallelism between the two versions can be at least partly traced to the problems of translating the original English items into Spanish.

Since languages are culturally endowed symbolic systems that do not represent the world in the same way (Sapir, 1921; Whorf, 1956), a test cannot be readily translated without compromising its reliability and validity. In conducting research on the English and Chinese versions of the major test used to screen kindergarten children in New York City,[14] Chu (1993) discovered that American and Chinese children often responded quite differently to what appeared to be the same tasks. On one task that required children to match a verbal description with three shapes concentrically arranged (i.e., one within another), many Chinese children interpreted the word *zhang* "middle" as indicating "center" rather than "between" and selected the innermost shape (rather than the shape located "between" the other two, which the test makers had set up as the key). In interpreting *zhang* in this way, they were following a well-established cultural pattern of working from the outer to the inner (e.g., when the Chinese give an address, they follow the order of country, state, city, street address; and when they give a date, they follow the order of year, month, day).[15]

As modern linguistics has taught us, cultural frames powerfully operate even in the most local details of language. Hill and Larsen (2000) provide many examples of African American children being misled by their cultural framing of a particular word or phrase as they deal with test material. One particular question on a reading test, for example, required that children interpret *home* as referring to a single place. When interviewing African American children who had difficulty with this question, the researchers discovered that many of them viewed *home* as referring to multiple places. As one 10-year-old African American girl put it, "I have a lot of homes—my grandmother's, my mother's, my aunt's, my cousin's, my grandfather's house" (p. 64). Having discovered this cultural

[13] The most extensive use of translation in large-scale testing is in Israel, where the admission test to higher education is offered not only in Hebrew but also in Arabic, English, French, Russian, and Spanish (see Beller, 1995, and Beller, Gafni, & Hanani, in press, for discussion of the psychometric properties of this test in its multiple language versions).

[14] This test is administered in English and four translated versions: Spanish, Haitian Kréyol, Mandarin, and Cantonese.

[15] Even Chinese rhetoric reflects this movement from the outer to the inner: in discussing social order, for example, classical Confucian texts first discuss the state, then the family, and finally the individual. This dominant cultural pattern is also evidenced in the decreasingly smaller boxes that are placed one within another, which, not surprisingly, bear the name "Chinese boxes."

point of view through interviewing, the researchers conducted an experiment confirming that African American children are, in fact, more likely than European American children to interpret *home* as referring to multiple places (even though experience with divorce or a country home led certain children in the latter group to make such an interpretation). Unfortunately this kind of subtle cultural difference is not readily picked up when test makers review prospective test material for cultural bias. It is for this reason that the College Board should consider using interviews and think-aloud protocols with culturally diverse students during the development of test material.

Policy recommendations

Before making specific recommendations, I would like to establish a general framework in which to approach accommodations. The use of accommodations is likely to increase in many state and federal testing programs, since legislation has been passed that mandates this approach to assessing culturally diverse students and students with disabilities. As long as these accommodations are used in tests such as NAEP, they will not be controversial. Since the purpose of such tests is to report on the state of American education, any news that these students, by virtue of the use of accommodations, are performing better will be welcomed.

If accommodations were, however, to be introduced into the SAT, there would be a public outcry that culturally diverse students are being given an unfair advantage.[16] In the case of first-language or bilingual accommodations, such protest would be misguided, since the research suggests that such accommodations can inhibit the performance of these students. As Abedi, Lord, and Hofstetter (1998) point out, culturally diverse students who have been instructed in English perform less well when they are tested in their first language.

Unless our educational system provides more opportunity for students to develop academic knowledge and skills in their first language, it is likely to perpetuate the condition of diglossia (Ferguson, 1957), in which bilingual individuals develop different competencies in different languages. Many Spanish-speaking students within American society, for example, use Spanish for conducting interpersonal relations among family and friends and English for processing technical information in school and in the larger society. Despite these complementary roles for the two languages, Spanish-speaking students are inclined, when faced with a stressful testing situation, to draw on their more familiar

[16] Such public reaction took place when the media reported on the "strivers" research at the Educational Testing Service. This research is concerned with developing a system that would identify as a "striver" any student who scores 200 points above the average score of students with a similar background (which can include variables such as family income, parents' education, or race and ethnicity). Prominent newspapers such as the *Wall Street Journal* gave the impression that this system was actually in place, which prompted a considerable protest from various organizations such as the Center for Individual Rights, a Washington-based nonprofit law group that has sued colleges over affirmative-action policies. In response to this controversy, Gaston Caperton, the president of the College Board, wrote a letter to college and university administrators in order to assure them that the research was only exploratory and there were no plans to put a system in place.

language. They are accustomed to conducting informal mathematical operations in Spanish, which may be one reason they are tempted to draw on it in a testing situation (see Lave, 1988, for a discussion of mathematical reasoning in everyday life). They are not, however, accustomed to using Spanish to carry out formal mathematical operations and so tend to perform less well.

Added to these problems are those associated with the bilingual processing of information. Since languages do not represent the world in the same way, a whole host of metacognitive and metalinguistic problems arise whenever information is simultaneously processed in more than one language. In a testing situation, the resolution of such problems can take up a good deal of time (not to mention the additional time required to process the content in two languages). It is thus not in the interests of culturally diverse students to respond bilingually in any testing situation that requires them to work rapidly (in certain NAEP accommodations, additional time was allotted, but if this policy were followed in a high-stakes test such as the SAT, it would meet with the same resistance discussed earlier).

Given all these problems, the College Board should not develop accommodations for culturally diverse students based on either a first-language or bilingual version of the SAT. It should consider, however, the possibility of strategically using the Spanish-language tests developed by its Puerto Rico office with Hispanic/Latino students. This office has wisely avoided any translation of tests used in this country. Rather it has opted to develop tests culturally appropriate for use in Latin American countries. The *Prueba de Aptitud Académica*® (PAA™) includes two reasoning tests—verbal and math—and three achievement tests (Spanish, English, and mathematics). These tests are now used by 127 institutions of higher education in these countries as well as in Spain. Within this country, the PAA could be used with certain Hispanic/Latino students, especially those who are recent immigrants, to determine whether they possess the requisite academic knowledge and skills to perform successfully at the college level, even though their English skills may not yet be sufficiently developed. As immigration from Latin American countries continues to increase, the PAA could be used to complement the SAT and thus play a strategic role in facilitating more informed admission decisions. If these tests are used, it is important that they be used judiciously. Manuel Maldonado-Rivera, the director of the Puerto Rico office, pointed out at the Roundtable that the PAA would be unduly difficult for Hispanic/Latino students who have not had a strong academic preparation within a Spanish-speaking country.

As for the use of modified English, the College Board should adopt this approach not as a test accommodation for culturally diverse students, but as a general policy that governs its test development for all students. Given that the College Board administers tests that have such high-stake consequences within American society, it needs to provide a leadership role in ensuring that language complexity not dictated by content be avoided

as a means of discriminating among students. As the research of Abedi, Lord, and Hofstetter (1998) indicates, whenever arbitrary complexity is avoided, the performance of all test takers improves, but it is the performance of culturally diverse students that improves most significantly. In order to deal with this issue, the College Board, in consultation with CRESST, could establish the review committee recommended in the preceding section.

Long-term Development of New Approaches

In this section I first explore how new technologies can be used to transform assessment practices in ways that can benefit culturally diverse students. I then discuss ways in which these students' potential for learning and sociolinguistic repertoires can be appropriately assessed.

Use of New Technologies in Assessment

There has been a reluctance to introduce technology into assessment for fear that it will only widen the gap between culturally diverse students and European American students. This concern is well founded, since school districts in which the majority of students are African American or Hispanic/Latino American generally have fewer technological resources available and fewer teachers prepared to make appropriate use of technology. In observing these students interact with technology in New York City schools, however, I have been struck by the ways in which their creative and critical faculties are often more fully engaged when print is augmented by speech and visual imagery. At the same time that educational policymakers deal with the equitable distribution of technology, we need to move forward in exploring how it can be used to transform assessment in ways that might benefit culturally diverse students.

Young people use a wide range of media to gain information about the world—magazines, books, radio, television, film, and the new computer media that increasingly bring together speech, writing, and various kinds of visual imagery. If educational assessment is to become more relevant to these students, it needs to move beyond its traditional reliance on isolated print. Rather than focusing on a narrow set of technical skills (e.g., extracting factual information from written language), assessment needs to deal with how effectively students can integrate and critically evaluate information from a range of sources such as radio, TV, film, and computers. Indeed, as we move toward such assessment, an Internet with well-developed multimedia capacities—or even a more robust successor we cannot yet envision—may have become the major source of information for young people in our society.

The Pacesetter program developed by the College Board encourages college-bound students to explore not only a broad range of media but also cultural diversity. Consider, for example, the curriculum of its English course, *Voices of Modern Cultures*:

Unit 1: Introduction to *Voices of Modern Cultures*

Unit 2: "Stranger in the Village": Encountering the Other, Being the Other

Unit 3: Culture and Voices in a Single Text

Unit 4: Inheriting Earlier Voices

Unit 5: Film, Language, and Culture

Unit 6: Mediating Culture: The Representation of Events and People

The emphasis on cultural diversity within the curriculum provides promising material around which to develop technology-based assessment. For example, James Baldwin's essay "Stranger in a Village" in Unit 2 could be presented in a hypermedia format with links to (1) intertextual allusions within the essay (e.g., Ellison's *Invisible Man*), and (2) the sociocultural context in which the essay was written (a Swiss village in winter). Students would explore these hyperlinks in order to respond to tasks dealing with the context—textual as well as cultural—of the essay.

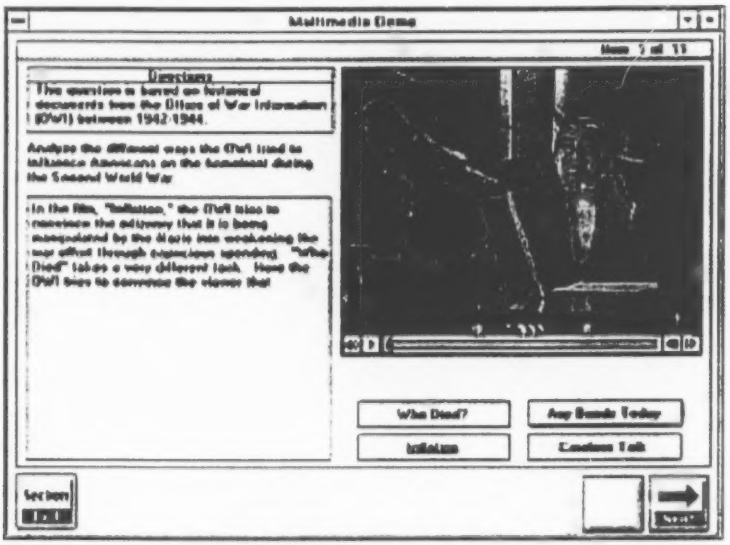
Unit 6 requires students to deal with "several media covering the same event [e.g., newspaper, radio, television, Internet] or different representations [of the same event] within a single medium [e.g., different television networks covering the same event]" (College Board, 1996, p. 13). The computer could be used to present contrasting video clips of television network coverage of an important event, and students would be asked to analyze differences in point of view.

Bennett (1998) explores such differences within the context of the AP Exam in United States History, which requires students "to use artifacts (e.g., excerpts from diaries, news articles, letters, maps, and political cartoons) as evidence to formulate a written argument in response to a question prompt" (p. 6). He then provides an example (see Figure 2), which demonstrates that "computers can accommodate all the artifacts used on the conventional test as well as historical films, TV, and radio, thereby extending the range of source material with which students must be proficient" (pp. 6–7).

In order to carry out this task, students must access four clips from films produced by the Office of War Information (OWI) during World War II by clicking on the four boxes in the lower right corner of the screen. After viewing the clips, they analyze the different ways in which the OWI tried to influence Americans during the war.

One potential advantage of such a task for culturally diverse students is that it diversifies the sources of information that they work with in a testing situation. Rather than being forced to rely only on written language, they have the opportunity to work with speech and visual imagery. It may well be that this interaction of visual imagery, oral language, and

FIGURE 2 Multimedia Question for AP Exam in United States History



(Bennett, 1998, p. 6. Reprinted by permission of Educational Testing Service.)

written language can stimulate these students to draw more effectively on the knowledge and skills that they possess.

Other kinds of technology-based tasks are now emerging that could be adapted for use in this course. To take just one example, in a project that Columbia University has developed with Chinese universities, students are required to (1) use the Internet to search for information on a particular topic, (2) synthesize the diverse bodies of information they find, and (3) communicate the results in written form. The integrated activities of this task—searching, synthesizing, communicating—reflect what students must do to perform successfully not only in higher education but also in a modern workplace.

In closing this section, I would like to challenge the College Board to provide national leadership in integrating the new technologies with educational assessment (see Everson, 1999, for a discussion of various ways in which such integration might be achieved). As president of the College Board, Gaston Caperton has been especially attentive to the potential of the new technologies. At the same time, he has been working closely with Edmund Gordon on issues of cultural diversity. It is thus a strategic time for the College Board to explore how the new technologies can be used to transform educational assessment in ways that would benefit students from a range of cultural backgrounds.

The Office of Educational Research and Improvement in the United States Department of Education has recently provided funding to set up the Center for Assessment, Culture,

and Technology to be jointly housed at the College Board and Teachers College, Columbia University. The initial project of the Center will be to develop and evaluate technology-based assessment tasks for the Pacesetter English course. A central goal of the project is to compare how students from a range of cultural backgrounds perform on the new tasks and on the SAT. Such a comparison will provide useful information not only on cultural differences in student performance on the two kinds of assessment but also on the quality of feedback for individual students that each provides. In the years ahead, the Center will be conducting a range of research projects that can shed light on the ways in which the new technologies can be used to transform assessment so that it better serves students from a range of cultural backgrounds.

Assessment of Potential for Learning

Our system of educational funding leads to a chronic shortfall in schools where culturally diverse students are heavily enrolled. Such lack of funding leads, in turn, to diminished academic achievement. Hence it becomes important to explore ways of assessing the learning potential of students whose lower achievement is often due to poverty and inferior schooling.

The wide range of research and practice on assessing students' potential for learning can be traced to the seminal work by Vygotsky (1962, 1978) that stressed the intimate relations between cognition and social interaction. Within the former Soviet Union, there evolved a widely used framework that is often referred to as *dynamic assessment* (e.g., Gal'perin, 1966; Ginzburg, 1981; Goncharova, 1990; Vlasova, 1972) which then spread to various European countries such as Germany (e.g., Guthke, 1992, 1993; Guthke, Beckman, & Dobat, 1997) and the Netherlands (e.g., Hessels & Hamers, 1993). At the same time, in Israel Feuerstein was developing his own approach to dynamic assessment which focused on what he calls cognitive modifiability (e.g., Feuerstein & Krasilowsky, 1972; Feuerstein & Rand, 1974). In recent years, the approach of Feuerstein and his colleagues has had increasing influence among American educators (e.g., Costa, 1999; Presseisen, 1999). Within this country, Budoff (1987a, 1987b) and Campione and Brown (1979, 1987) have also been prominent in the use of dynamic assessment.

Dynamic assessment differs most markedly from traditional testing in that individuals are provided additional information when they are unable to perform a certain task. The assessor then attends to how well they use this information in resolving their difficulties. Dynamic assessment also differs from traditional testing in its greater attention to the processes that underlie the performance of tasks. Hence individuals are encouraged to explain what they are thinking as they perform. As Grigorenko and Sternberg (1999) point out, dynamic assessment is more congruent with what we do when we teach; as they put it, "The conventional attitude of neutrality is thus replaced by an atmosphere of teaching

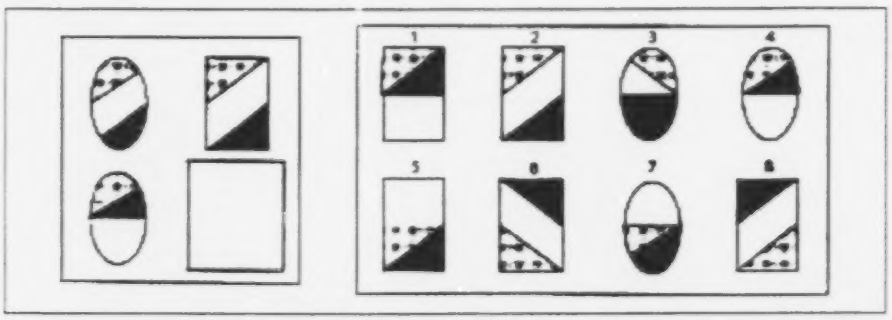
and helping" (p. 75). Given such a supportive atmosphere, dynamic assessment holds out a good deal of promise for assessing culturally diverse students who often approach a testing situation with considerable apprehension about how they will perform (Steele, 1997).

Nevertheless, dynamic assessment faces major problems that need to be resolved before it can be widely used. First of all, it is difficult to standardize the feedback to be provided an individual who makes an inadequate response. In the case of a multiple-choice task, such standardization is easier to carry out, for if research is conducted on why students choose the various distractors, then feedback can be tailored to each of those choices. Yet those who work with dynamic assessment tend to prefer constructed-response tasks as a means of gauging learning potential. The range of human responses to such tasks, however, is often so great that it is difficult to provide standardized feedback. If tasks are constructed that elicit a fairly restricted range of responses, standardized feedback can be more readily provided (see Wang & Hill, 1999, for research on constructing such tasks).

Here, too, new technologies may have an important role to play. If tasks are administered online, the computer can be programmed to provide standardized feedback according to the response an individual makes. If, for example, a student response leaves out a crucial component, a prompt can be provided that focuses on this component. If the student response to this feedback is appropriate, the program can then stop. If, however, the response is not appropriate, the program can provide a more detailed prompt. Such branching networks are at the heart of computer programming and have considerable potential in standardizing feedback within dynamic assessment.

Added to the problem of standardizing feedback is that of providing ecologically valid tasks. Dynamic assessment has been built largely around a repertoire of specialized tasks such as Raven's progressive matrices that require students to construct patterns as they work with geometric shapes. Figure 3 shows one of these tasks used within the Learning Potential Assessment Device (LPAD).

FIGURE 3 A Matrix Task from the LPAD



(Feuerstein, Falik, & Feuerstein, 1998, p. 124. Reprinted with permission of the authors.)

Students are required to select the first shape from the 8 shapes in the box on the right in order to construct the appropriate pattern in the box on the left. Such specialized tasks are designed to minimize not only the role of language but also prior academic achievement. Yet the fact that they do not make use of familiar activities can cause confusion, which is likely to be even greater for culturally diverse students who often have had less exposure to such activities. In an earlier section (see p. 27), we considered an example of how Chinese children were led astray on a task used in New York City to assess learning potential. Although the confusion was partly related to the translation of the word *middle*, it was also related to a Chinese preference for structuring information from the outer to the inner.

There are no easy solutions to such problems, though it may be useful to consider how they have been handled in an assessment program that teachers and I have developed for a metropolitan New York school district where the population is culturally diverse (Hill, 1999b). The program is called *The Progress Profile* because each student's performance is evaluated in relation to prior performance rather than to a predetermined standard or to the performance of other students (i.e., criterion-based or norm-based testing). The program also incorporates the two basic features of dynamic assessment: documenting the processes that students engage in and providing feedback when their response to a task is in some way lacking.

This program does not present specialized tasks but rather tasks that students are familiar with, since the teachers and I are committed to the notion of ecological validity: that is to say, students are better able to show what they know and can do when they are working with familiar routines. Hence in assessing their reading comprehension, they are asked to retell a story that they have read and then respond to tasks that probe factual, inferential, and experiential aspects of their comprehension (Bloom, 1984). We have found that this activity is particularly appropriate for culturally diverse students, since they often engage in story retelling in their homes and communities as well as in school (the stories that we use are drawn from a range of cultural traditions).

If students falter during the story retelling, the assessor can provide a prompt to determine whether they can use it effectively in moving forward; or if they fail to make a certain inference, the assessor can provide details from the story to see whether they can put them together in order to make an appropriate inference. Since the retelling of a story is a familiar task, students are not confused about what they are expected to do. The responses they make can shed a good deal of light on their capacities to deal with story structure: for example, the degree to which they attend to the spatial setting and the temporal sequence of events; the degree to which they attend to the plot and its supporting details; and the degree to which they attend to character motivation. In analyzing how students retell the story, we have found that a basic distinction in styles—which we refer to as *summary* and *performance*—can help us understand the cognitive styles that they use

in approaching other tasks (see Hill & Larsen, 2000, p. 378 for examples of these two styles of retelling a story).

The College Board might consider working with researchers such as Robert Sternberg and Joseph Campione to establish a working group to explore how dynamic assessment might be best used to evaluate the learning potential of culturally diverse students. Appropriate assessment of these students' potential for learning could have practical applications for the College Board as it deals with students in its college preparatory programs. In the case of students who have experienced poverty and inferior schooling, an assessment of potential for learning could be useful in making decisions about the most strategic ways to work with them. Such assessment could also be used by institutions of higher education not only in deciding whether to admit certain students but also, once they are admitted, in determining how to provide effective support.

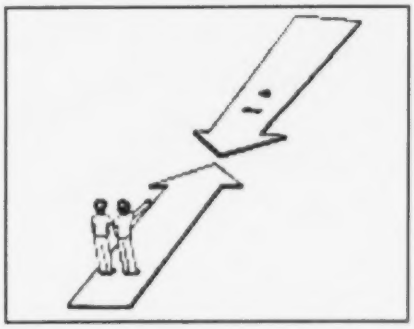
Assessment of Sociolinguistic Repertoires

Many culturally diverse students develop complex sociolinguistic repertoires in responding to the demands of the various worlds in which they live. Such repertoires represent strengths that have gone virtually unexplored in educational assessment (Valdes & Figueroa, 1994). Not only have these strengths been ignored, they have often worked against these students when they are faced with a test such as the SAT in which they must work with the features earlier described as speededness, arbitrary language complexity, and the multiple-choice format. By its very nature, a complex sociolinguistic repertoire introduces more options to consider and thus leads to more processing time, which can be counterproductive when working under time pressure.

Assessing these repertoires is a challenging task, and the College Board needs to assemble expert help in order to carry it out properly. Although a good deal of research has been conducted on such repertoires, it has often focused on surface features of language performance—phonological and syntactic—at the expense of underlying frames that have to do with what linguists describe as semantic and pragmatic aspects of performance.

To give just one example of research dealing with these deeper aspects, doctoral students and I have investigated distinctive ways in which African American students use spatial and temporal frames, even when they appear to be speaking standard English (Hill, 1991a, 1991b; McKenna, 1985). To illustrate one of these ways, let us consider what appears to be a relatively straightforward use of the words *front* and *back* to describe the relation between two objects—for example, a ball and a pen—that are not thought of as having a front and back. Most European American speakers of standard English view these objects as being oriented toward themselves when they speak. Figure 4 illustrates this perspective, which is described with the term *mirror-image*.

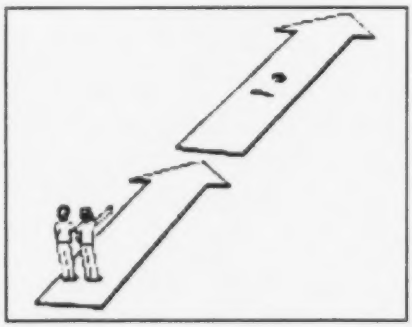
Figure 4 A Mirror-Image Perspective



"Hey, there's my pen in front of the ball."

Many African American speakers, however, view these objects as oriented away from themselves when they speak. Figure 5 illustrates this perspective, which is described with the term *in-tandem*.

Figure 5 An In-Tandem Perspective



"Hey, there's my pen behind the ball."

As Mooney and Goldstein (1982) have shown, African American speakers can shift from an in-tandem perspective to a mirror-image perspective when they shift from one sociolinguistic setting to another (i.e., from communicating with an African American peer to communicating with a European American teacher). Such shifting is fundamental to what culturally diverse students do as they use language in varied sociocultural settings. Hence any reliable assessment of their repertoires must be sensitive to the setting in which they are using language.

In extending this research to Spanish-speaking students in Puerto Rico, Reyes (1994) introduced an activity that could be useful in assessing culturally diverse students'

potential for cognitive modifiability. This activity was originally developed by McKenna (1985) as a way for researchers to deduce whether a student uses a mirror-image or an in-tandem perspective. The researcher and the student sit side by side at a table on which two playing cards (with different numbers) are placed face down (one card is placed nearer than the other to the two players). The researcher then explains that the purpose of the game is to try to guess which card has the higher number. The student is then asked whether (s)he would like to choose the card that is "in front of the other one" or the card that is "behind the other one." After responding with either "in front" or "behind," the student is asked to turn over the card(s) he has selected. The researcher is then in a position to deduce which kind of response the student has made: if, for example, the student selected "the one in front" and then turned over the further card, (s)he is using an in-tandem perspective.

Reyes extended the use of the card game in a novel way. Working with Puerto Rican high school students, she first used the game to identify whether individual students use a mirror-image or an in-tandem perspective. She then paired students who use opposing perspectives and asked them to play the game together. During the first stage of her research, Reyes found that students were almost evenly divided between the two kinds of responses (she traces this split to the island's complex heritage, which blends African and European cultures).[17] During the second stage, she discovered that more students shifted to an in-tandem perspective than to a mirror-image one (thereby uncovering evidence that this perspective is stronger in the peer culture). She also discovered that students, while playing the card game, negotiate their differing spatial points of view in quite different ways. Certain negotiations were dismissive in that students accused each other of cheating or being stupid (e.g., *Chacho, no seas tan sángano* "Man, don't be so stupid"). In other negotiations, however, students showed a capacity not only to become aware of the contrasting point of view but even to adapt to it (e.g., *Sí, yo creo que ella lo está viendo así como si fuera una fila y la de atrás es la que sigue* "Yes, I think that she is seeing it as if it were a row and that the one in back is the one that follows").[18]

Although this research was conducted in two stages and is thus not easy to administer, it exemplifies the kind of activity that might be used in the dynamic assessment of culturally diverse students' potential for learning. To begin with, this activity reflects Vygotsky's notion that cognition is socially situated: this activity has the particular merit that the fundamental social situation is comprised of two students rather than a student and an assessor. This activity has the further merit that it uncovers students' capacity for

[17] This spatial and temporal research has been conducted in many countries around the world and shows that underlying frames are often transmitted as people move from one country to another. It has shown, for example, not only African and African American continuities (Abubakar, 1985; Hill, 1974, 1975a, 1975b, 1978, 1982; Isma'il, 1979, Mahmoud, 1997) but also Chinese and Chinese American continuities (Hill, 1991a; Ho, 1998; Ji, 1998; Wei, 1994) as well as Korean and Korean American continuities (Kim, 1996).

[18] Unlike the first example, where the student is directly addressing his partner, the student in this example is addressing the researcher. It is as if she has become aware of the underlying difference and wants to objectify it.

cognitive modifiability while they are at play with a partner. As the vast potential of the new technologies is tapped, similar kinds of activities might be developed that will allow us to uncover students' potential for learning while they are engaged in familiar kinds of interaction.

In closing this section, I would like to return to culturally diverse students' sociolinguistic repertoires and provide an example of how technology might be used to explore these repertoires in a bilingual setting. Using the multimedia approach displayed in Figure 2 (see p. 32), film clips could be presented in which English and another language are used together in certain scenarios: for example, a courtroom where an immigrant is forced to rely on an interpreter. Students could engage in various activities that involve transcription, translation, and interpretation. These activities should be designed to show that they understand not only the two languages but also the ways in which they are strategically interfaced (e.g., how the interpreter draws on colloquial language to translate what the judge expresses in technical language). To conduct such assessment is obviously challenging, for as Harris (2000) points out, "There is more language, not less language, to consider in the bilingual assessment process" (p. 147).

If the languages that culturally diverse students bring to school are developed in concert with English, they can provide the bilingual resources that are increasingly needed for effective global communication. It should thus be a matter of national policy to insure that these languages are properly developed within our educational system. As Baugh (1995) has put it, "For too long our education policies have squandered valuable linguistic resources under a shroud of linguistic chauvinism that is in direct opposition of our need to participate fully in the global economy, where knowledge of other languages and cultures is most desirable" (p. 298).

A Further Research Agenda for Culturally Diverse Students

As the College Board develops a strategic plan for dealing with culturally diverse students, it will need to conduct research that extends beyond developing new approaches to assessment. In laying out such a research agenda, I first focus on ways in which the College Board might use existing data not only to sharpen its characterization of these students but also to communicate more effectively with the larger public. I then present ways various research projects that can shed new light on these students themselves.

More Effective Use of Extant Databases

In American society, public discourse tends to reduce the complexity that can be found within various culturally diverse populations. In order to counter this tendency,

institutions such as the College Board need to provide leadership in characterizing the diversity not only across various groups but also within each of them. In administering its programs, the College Board collects substantial bodies of information that can be used to provide more refined characterizations of students within each group. As a good example of such characterization, we have seen how the College Board specifies two countries of origin—Mexico and Puerto Rico—in characterizing the Hispanic/Latino students who take the SAT and AP exams.

The College Board collects the following kinds of information about the students taking these exams:

1) gender
2) race/ethnicity
3) country of origin
4) home language(s)
5) social class
   (a) family income
   (b) parental education

These different kinds of information can be used to provide more refined characterizations of culturally diverse students. In order to illustrate the problems with current reporting, Scott Miller makes the following point in a memorandum: "We don't have SAT score data readily at hand comparing (1) Mexican Americans who speak Spanish at home and whose parents have completed only grade school, and (2) Korean Americans who speak Korean at home and whose parents have completed only grade school. This means that we don't have studies looking for overprediction or underprediction patterns for these and other such subpopulations" (see Miller, 1995, for further discussion of the need for moving beyond broad categories).

In order to use existing data more effectively, the College Board might consider collaborating with the National Center for Education Statistics (NCES). It would be probably advisable to begin with a small pilot project that could be used to determine the feasibility of such a relation (and whether it leads to any undesirable federal intrusion into the operations of the College Board). One possibility would be to create databases on SAT performance for various subpopulations of culturally diverse students that could be used to explore the predictive validity of the SAT for these students.

New Research Projects

As indicated earlier, two controversial areas of research have to do with (1) the predictive validity of the SAT, and (2) the degree to which coaching can improve test performance. I here discuss research projects that could be conducted with culturally diverse students in

these two areas. I also discuss a project that reflects a priority of the National Task Force on Minority High Achievement: documenting factors that contribute to the academic success of these students who perform beyond expectations.

Predictive Validity of the SAT

The publication of *The Shape of the River* (1998) by William Bowen, the former president of Princeton University, and Derek Bok, the former president of Harvard University, has renewed public interest in the predictive validity of the SAT. Based on a study of more than 45,000 students who attended selective colleges, Bowen and Bok report that "rises in SAT scores have a modest correlation to increases in class rank, a modest correlation to the odds of going to graduate school, a modest correlation to the likelihood of graduation, and a modest relationship to future earnings" (cited in Schwartz, 1999, p. 51). More recently, this point of view has been reinforced by *The Big Test: The Secret History of the American Meritocracy* (1999a) by Nicholas Lemann, which tells the story of how "a test that predicts [only] about 15 percent of the variance in freshman grades in college [nevertheless] became a national obsession" (1999b, p. 51). These books written for an educated public bring the predictive validity of the SAT under increasing scrutiny.

The College Board might consider initiating new studies of predictive validity for the increasing number of culturally diverse students who are taking the SAT. These studies need to be designed with greater attention to various subpopulations of these students (especially to immigrant populations from Caribbean, Latin American, and Asian countries). The results of these studies will need to be calibrated with those reported in Ramist, Lewis, and McCamley-Jenkins (1994), where predictive validity of the earlier version of the SAT was reported for various subpopulations based on gender, ethnic identity, and language background (see Heubert & Hauser, 1999, and Jencks & Phillips, 1998, for further discussion of these issues).

One promising direction for new research would be to work closely with the Task Force on Minority High Achievement in investigating student performance within college programs specifically designed to support their high achievement. As Fleming and Garcia (1998) point out, the predictive validity of the SAT for African American students varies considerably according to the nature of the institution they attend. Although the performance of these students is often overpredicted at colleges and universities that have historically served European American students, it is usually underpredicted at institutions that have set up supportive programs for these students. Studies of predictive validity with respect to culturally diverse students will be more useful if they can factor in the degree of institutional support provided them.

In a report to the College Board, Gándara and Maxwell-Jolly (1998) explore issues of ethnocultural identity, gender, and social class in relation to programs designed to support

culturally diverse students in higher education. In proposing a research agenda, the authors call for greater attention to which programs work best for particular groups (they point out that "programs that appear to be effective for African Americans may not hold the same promise for some Hispanic groups," p. 93). They also argue for increased attention to the factors that contribute to the limited participation of male students within these programs (men receive only about one-third of higher education degrees earned by African Americans). Finally, they call for sharper focus on the degree to which these programs serve the needs of middle-class students. As Bowen and Bok (1998) point out, African American and Hispanic/Latino American students from a middle-class background reflect a significantly lower academic performance than their European American counterparts in colleges and universities that have historically served a largely European American population.

In conducting this more refined research, the College Board could be especially attentive to programs that have a strong track record in serving culturally diverse students. Among these are the Biomed Honors Corps at Xavier University, the Professional Development Program at the University of California, Berkeley, and the Meyerhoff Scholars Program at the University of Maryland, Baltimore County. Particular attention should be given to distinctive features of these programs that contribute to the outstanding academic achievement of the students that they serve. All these programs, for example, are characterized by both an insistence on high standards and a strong system of support for students. In a recent interview, William Massey, a research mathematician at Bell Labs, stressed these two qualities in speaking of the Meyerhoff Scholars Program (which often places student interns at the Bell Labs): "There are institutions that stress the achievement of excellence, but there's very little nurturing or encouragement to help students achieve that level of excellence; and there are programs that have nurturing, but they don't stress achievement of excellence. This is one of the rare programs that stresses both" (Schemo, 2000, A12).

Coaching for the SAT

As reported in a *New York Times Magazine* article (Schwartz, 1999), representatives of coaching programs make strong claims for improving student scores on the SAT. For example, Arun Alagappan, whose program Advantage Testing has been popular in metropolitan New York, states that "the mean gain for the students I tutored in 1997 and 1998 was 266 points, from the first time they took an official PSAT/NMSQT™ or SAT to the last. He further states that "over the years 91 percent of the students who worked with any of the New York City-based tutors at Advantage increased their score by at least 100 points" (pp. 31–32).

The article then discusses the College Board-commissioned study by Powers and Rock (1998), which, in addition to comparing the initial and final scores of students who had

received out-of-school coaching, compared their gains to those made by uncoached students (even in the absence of coaching, students generally improve when they take a test for the second time). This study found that the gains attributable to out-of-school coaching were much more modest than those advertised by major coaching schools (see Table 8).

TABLE 8

Gains on the SAT by Coached and Uncoached Students

Group	Initial Score	Final Score	Gain
Verbal			
Coached ($n = 427$)	500	529	29
Uncoached ($n = 2,733$)	506	527	21
Math			
Coached ($n = 427$)	521	561	40
Uncoached ($n = 2,733$)	505	527	22

Note: Adapted from Powers & Rock, 1998, p. 7.

There are several drawbacks to the Power and Rock study. To begin with, although they are investigating the "effects of coaching," they define coaching quite narrowly as "any and all activities conducted in special preparation programs offered to students outside their schools" (p. 1). This approach seems to include mainly privately paid for programs and to exclude coaching offered as part of regular classes or the special programs which are increasingly offered by schools and which may rival those conducted by private coaching services.

Another drawback is that the sample used in the study was largely European American or Asian American. As shown in Table 9, 79 percent of the coached sample is made up of these two groups, while the two groups focused on in this report constituted only 17 percent (among the 427 students, 47 were African American and 26 were Hispanic/Latino American).These numbers are so small that they call into question the statistical reliability of the study with respect to these groups. The underrepresentation of these two groups was unavoidable, since the coached sample was built around students whose families could afford to pay for private coaching: 43 percent of the coached students, but only 18 percent of the uncoached students came from families whose annual income exceeded $80,000. In addition, the coached students had higher grades, had taken more math and science courses, and had more test-taking experience.

TABLE 9

Distribution of Various Groups in the Coached and Uncoached Samples

Group	Coached	Uncoached
African American	11%	9%
Asian American	21%	8%
Hispanic/Latino American	6%	8%
Native American	1%	1%
European American	58%	72%
Other	4%	3%

Note: Adapted from Powers & Rock, 1998, p. 8.

Hitherto no large-scale study has been conducted as to how students who do not have this kind of background respond to coaching. It may well be that coaching would enable them to gain far more than students whose home and school experience has already imparted a good deal of the knowledge and skills that are transmitted in a coaching program. The College Board should consider undertaking research on coaching such students. Since these students would not be in a position to pay for commercial coaching programs, the College Board should explore providing free services to these students as part of its research and development of effective test preparation materials to be delivered online. One of the major goals in developing these materials is to provide high-quality, low-cost test preparation to the increasing numbers of students who are seeking admission to higher education. Since many of these students are culturally diverse, the College Board could carry out research while it provides test preparation (it could also investigate how well online delivery of such preparation compares with a more traditional delivery).

In addition, the College Board could explore linking up with key states in providing test preparation. Many states are making plans to provide such preparation in programs that target these students (e.g., in California, the College Preparation Partnership Program is just now getting underway). If the College Board were to develop an effective working relation with a large state such as California, it would be in a position not only to provide test preparation but also to evaluate how effective it is with these students.

By undertaking such research, the College Board could strategically address criticism from various organizations. The NAACP has claimed that the widespread use of out-of-school coaching discriminates against culturally diverse students whose families cannot afford to pay for it and has thus begun a national effort not only "to reduce the influence of College Board scores in admissions, [but also] to work for state-financed test preparation for all minorities." As Jeffrey Johnson, the national coordinator of the NAACP Youth Council, puts it, "It's irresponsible to tell children to be successful but not

empower them with the tools they need" ("NAACP Seeks," 1999, p. A23). The College Board could work with the NAACP in providing high-quality coaching for students who cannot afford to pay for it.

Such an undertaking would also disarm critics such as Robert Schaeffer, public education director for FairTest, who has been particularly critical of the Powers and Rock study. As he explained during an interview, "It's like tobacco-company research. The College Board pays for it, they design the methodology, they ask the questions, they interpret the data and, lo and behold, they get the result they want" (Schwartz, 1999, p. 34). No matter how exaggerated such criticism may be, it is likely to continue unless the College Board adopts a bolder approach to research on coaching.

As Alagappan of Test Advantage has pointed out, the gains that students make through coaching do not invalidate the SAT. As he puts it, "I'm a student of standardized tests, and I believe that in a three-hour block of time, the SAT is the best one out there by a good margin" (Schwartz, 1999, p. 32). He claims that the SAT can be used "as a source of motivation for students to learn a range of skills that...are important in any intellectual endeavor—among them the capacity to break down and reason through a difficult problem." Or as he puts it more colloquially, "Try to give a kid several thousand words to learn without the SAT coming up....There's no way. You need a reason, and that's the beauty of a test" (p. 56).

In closing this section, I would like to call attention to a special group of culturally diverse students that the College Board could consider for a coaching program: those whom teachers in the Pacesetter and AP programs regard as academically strong, but who do not perform well when they are taking a high-stakes test.[19] Any coaching program developed for such students should take account of the research of Claude Steele and various colleagues. As they have shown, certain academically talented African American students perform significantly better when they engage with material apart from a testing situation (see, for example, Steele, 1997; Steele, 1999; Steele & Aronson, 1995). In accounting for their weaker performance when taking a high-stakes test, Steele traces it to the threat of the negative stereotype that society projects on to them (such a threat has also been hypothesized by Steele as explaining the poorer performance of women on high-stakes testing of mathematics).

In developing a coaching program for these African American students, special effort should be made, as Steele puts it, "to desensitize" them to the negative threat they associate with a testing situation. In attending to this social psychological dimension of student performance, it is important not to neglect the factors identified with testing itself that were earlier identified as problematical for these students: speededness, arbitrary language complexity, and the multiple-choice format. As long as these factors continue to play a

[19] As Stefanakis points out in *Whose Judgment Counts?* (1998), the professional judgment of teachers should be increasingly used in research.

prominent role in testing, it is important that test preparation programs focus on strategies for dealing with them.

Case Studies of Successful Students: Family, Community, and School

As the College Board seeks to strengthen its precollegiate programs, it might consider undertaking research on culturally diverse students who are performing at an unexpectedly high level within these programs. As the report of the National Task Force on Minority High Achievement (College Board, 1999) points out, so many national programs are directed at students who are underperforming that those who are successful are often ignored.

Thus far the research on students who perform beyond expectations has focused largely on their standardized test performance, as exemplified by the "strivers" research at the Educational Testing Service. Other researchers have conducted similar research but not with the goal of adjusting test scores. Rather their goal, as Michael Nettles of the University of Michigan puts it, "is to try to find out how those high achievers did it, so that more students in their conditions can do better as well" (cited in Gose, 1999, p. A55).

The College Board might also consider conducting research on unexpectedly high achievers from such an explanatory perspective. Using a case study methodology, it could focus on students in its precollegiate programs—for example, the Pacesetter program—who perform beyond expectations with respect to both test scores and course grades (in designing studies on the effects of coaching, the preceding section recommended that the College Board be attentive to students in these programs who maintain a high level of academic achievement but who perform poorly on tests such as the SAT).

Such case studies would be valuable for a number of reasons. To begin with, the methods used in most research sponsored by the College Board are quantitative. Although such methods yield useful information, they do not always provide explanatory insight into factors that contribute to academic success. Case studies, if properly carried out, can provide information that is useful in shaping policy and practice.

Moreover, a focus on culturally diverse students who are academically successful can help break down stereotypes. So much research deals with the problems associated with these students that the considerable strengths they draw on are often overlooked. These strengths need to be documented not only in individuals (e.g., work habits that lead to timely and thoughtful completion of assignments) but also in the homes they come from (e.g., the ways in which parents nurture effective work habits). It is also important to identify community resources that successful students draw on. According to the report of the National Task Force on Minority High Achievement (College Board, 1999), supplementary educational programs in the community are crucial to the success of these students (it calls particular attention to the effective ways in which East Asian communities have provided such support).

Case studies of these students should also identify factors in school settings that contribute to their success. Here, too, they derive strength from crucial personal relationships not only with teachers and administrators but also with fellow students. The report of the National Task Force on Minority High Achievement (1999) cites the important role that supportive peer groups often play in the success of these students.

Beyond these supportive relationships are the various kinds of educational practices that best serve culturally diverse students: the curricular materials and the instructional methods that engage their interest as well as the assessment models that elicit what they know and can do. It is particularly important to document educational uses of technology that are effective with these students: one highlight of the case studies could be to display projects in which students have effectively used technology in working with their own cultural backgrounds. These projects could be related either to the public domain (e.g., projects having to do with historical events) or to the personal domain (e.g., projects having to do with family history).

If such case studies are properly carried out, they could shed new light on factors crucial to the academic success of culturally diverse students. These studies could also be used to provide role models for promising middle school students and their families. The College Board could publish the most compelling case studies as a means of recruiting culturally diverse students: well-crafted stories of academic success can be far more attractive than tables and charts in showing how a program such as Pacesetter, in concert with the home and community, can nurture academic success for these students.

Conclusions

In concluding, I would like to provide an outline of the major recommendations that have been put forth:

1) The College Board should exercise national leadership in maintaining exemplary academic preparatory programs for culturally diverse students. In order to achieve this goal, it might explore different forms of collaboration with leading graduate schools of education that have been traditionally committed to preparing teachers to work with these students.
2) The College Board should develop appropriate testing and assessment for culturally diverse students.
   (a) As a matter of short-term policy, it might consider establishing a review committee whose responsibility would be to broaden the base of test development for the SAT. This committee would be particularly concerned with reviewing the language used on the test in order to insure that arbitrary complexity is not introduced as a means of discriminating among students.

(b) As a matter of long-term policy, it could explore ways in which
   (i) new technologies can be used in assessing a broader range of abilities;
   (ii) new assessment models can be developed to evaluate these students' potential for learning as well as their sociolinguistic repertoires.

3) The College Board should establish a research agenda that goes well beyond testing and assessment with respect to culturally diverse students.
   (a) It might redesign its extant data bases so that they can yield more refined information about these students (in carrying out this redesign, it could explore collaborating with the National Center for Educational Statistics).
   (b) It could explore new research approaches to two basic issues:
      (i) the degree to which the SAT accurately predicts the academic performance of culturally diverse students;
      (ii) the degree to which these students can benefit from coaching programs for the SAT.
   (c) It might consider developing case studies that focus on factors in the family, community, and school that contribute to the academic success of these students in its college preparatory programs.

As the College Board develops a set of strategic policies with respect to culturally diverse students, it will need to strengthen its alliances with organizations such as CRESST. Through these strategic alliances, the College Board can maintain leadership in responding to the growing challenge that higher education is facing: the number of culturally diverse students seeking admission to higher education is rapidly increasing, but their academic preparation continues to lag behind.

In closing this report, I would like to emphasize that our colleges and universities are strengthened as they achieve their goal of greater cultural diversity: not only is the diversity in the larger society more equitably represented, but it also functions as a crucial educational resource. Students learn more effectively when they are exploring new knowledge with others who do not share their own background. In negotiating various cultural worlds, they develop not only a greater capacity for understanding divergent points of view but also a greater range of communication skills that they can put to good use in the larger society.

The nation as a whole thus benefits from increased cultural diversity in higher education. As culturally diverse students pursue higher education, they will be in a position to draw on their language skills to facilitate international exchange in educational, cultural, and commercial spheres. If this nation can insure that its culturally diverse citizens have the opportunity to pursue high-quality education at advanced levels, it can make a far richer contribution to global society. These citizens have special knowledge and skills that neither our nation nor the world can afford to leave undeveloped.

References

Abedi, J., Lord, C., & Hofstetter, C. (1998). Impact of selected background variables on students' NAEP math performance (CSE Technical Report 478). Los Angeles: UCLA Center for the Study of Evaluation/National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., Lord, C., & Plummer, J. (1995). *Language background as a variable in NAEP mathematics performance: NAEP TRP Task 3D: Language background study.* Los Angeles: UCLA Center for the Study of Evaluation/National Center for Research on Evaluation, Standards, and Student Testing.

Abubakar, A. (1985). *The acquisition of 'front' and 'back' among Hausa children: A study in deixis.* Unpublished doctoral dissertation, Teachers College, Columbia University, New York.

Adames, J. A. (1987). *A study of the reading process of selected English as a second language college students.* Unpublished doctoral dissertation, Teachers College, Columbia University, New York.

Allen, K. (1988). *The development of a test of communicative competence for speakers of English as a second language in Zimbabwe.* Unpublished doctoral dissertation, Teachers College, Columbia University, New York.

Anderson, N., Jenkins, F., & Miller, K. (1996). *NAEP inclusion criteria and testing accommodations: Findings from the NAEP 1995 field test in mathematics.* Princeton, NJ: Educational Testing Service.

Anderson, N., & Olson, J. (1996). *Puerto Rico assessment of educational progress: 1994 PRAEP Technical Report.* Princeton, NJ: Educational Testing Service.

Aronson, E., & Farr, R. (1988). Issues in assessment. *Journal of Reading, 32,* 174–177.

August, D., & Hakuta, K. (1997). *Improving schooling for language-minority children: A research agenda.* Washington, DC.: National Academy Press.

Baugh, J. (1983). *Black street speech: Its history, structure, and survival.* Austin, TX: University of Texas Press.

Baugh, J. (1995). *Linguistics, education and the law: Educational reform for African American language minority students.* Linguistics and Education, 7, 87–105.

Beller, M., (1995). Translated versions of Israel's inter-university Psychometric Entrance Test (PET). In T. Oakland & R. K. Hambleton (Eds.), *International perspectives on academic assessment* (pp. 207–218). Boston: Kluwer.

Beller, M., Gafni, N., & Hanani, P. (in press). Constructing, adapting, and validating admissions tests in multiple languages. In R. K. Hambleton, P. Merenda, and

C. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment*. Hillsdale, NJ: Erlbaum.

Bennett, R. (1998). *Reinventing assessment: Speculations on the future of large-scale assessment*. Princeton, NJ: Educational Testing Service.

Bennett, R. (1999). Using new technology to improve assessment. *Educational Measurement: Issues and Practice, 18*, 5–12.

Bhasin, J. (1990). *The demands of main idea tasks in reading comprehension tests and the strategic responses of bilingual poor comprehenders*. Unpublished doctoral dissertation, Teachers College, Columbia University, New York.

Bhopal, R., & Donaldson, L. (1998). White, European, Western, Caucasian, or what? Inappropriate labeling in research on race, ethnicity, and health. *American Journal of Public Health, 88*, 1303–1306.

Bloom, B. (1984). Taxonomy of educational objectives. New York: Longman.

Bowen, W., & Bok, D. (1998). *The shape of the river: Long-term consequences of considering race in college and university admission*. Princeton, NJ: Princeton University Press.

Budoff, M. (1987a). Measures for assessing learning potential. In C. S. Lidz (Ed.), *Dynamic testing* (pp. 173–195). New York: Guilford Press.

Budoff, M. (1987b). The validity of learning potential. In C. S. Lidz (Ed.), *Dynamic testing* (pp. 52–81). New York: Guilford Press.

Butler, F., & Stevens, R. (1997). *Accommodation strategies for ELL large-scale assessments: Student characteristics and other considerations* (CSE Technical Report 448). Los Angeles,: UCLA Center for the Study of Evaluation/National Center for Research on Evaluation, Standards, and Student Testing.

Campione, J. C., & Brown, A. L. (1979). *Human intelligence*. Norwood, NJ: Ablex.

Campione, J. C., & Brown, A. L. (1987). Linking dynamic testing with school achievement. In C. S. Lidz (Ed.), *Dynamic testing* (pp. 83–115). New York: Guilford Press.

Carnevale, A., & Fry, R. (2000). *Crossing the great divide: Can we achieve equity when generation Y goes to college?* Princeton, NJ: Educational Testing Service.

Cheung, O., Clements, B., & Miu, Y. (1994). *The feasibility of collecting comparable national statistics about students with LEP*. Washington, DC. National Center for Education Statistics.

Chu, H. (1993). *Assessing Chinese kindergarten children in New York City*. Unpublished doctoral dissertation, Teachers College, Columbia University, New York.

College Entrance Examination Board. (1996). *Pacesetter English*. New York: Author.

College Entrance Examination Board. (1999). *Reaching the top: A report of the National Task Force on Minority High Achievement*. New York: Author.

Collier, V. (1989). How long? A synthesis of research on academic achievement in second language. *TESOL Quarterly*, *23*, 509–531.

Costa, A. L. (Ed.) (1999). *Teaching for intelligence II*. Arlington Heights, IL: Skylight.

Coyle, M. (1992). *The New Jersey high school proficiency test in writing: A pragmatic face on an autonomous model*. Unpublished doctoral dissertation, Teachers College, Columbia University, New York.

Cummins, J. (1981). The role of primary language development in promoting educational success for language minority students. In *Schooling and language minority students* (pp. 3–49). Sacramento, CA: California Department of Education.

Everson, H. (1999). A theory-based framework for future college admissions tests. In S. Messick (Ed.), *Assessment in higher education: Issues of access, student development, and public policy* (pp. 113–132). Hillsdale, NJ: Erlbaum.

Ferguson, C. (1957). Diglossia. *Word*, *15*, 325–340.

Feuerstein, R., Falik, L., & Feuerstein, R. (1999). In R. Samuda (Ed.), *Advances in cross-cultural assessment* (pp. 100–161). Thousand Oaks, CA: Sage Publications.

Feuerstein, R., & Krasilowsky, D. (1972). Interventional strategies for the significant modification of cognitive functioning in the disadvantaged adolescent. *Journal of the American Academy of Child Psychiatry*, *11*, 572–582.

Feuerstein, R., & Rand, Y. (1974). Mediated learning experiences: An outline of proximal etiology for differential development of cognitive functions. *International Understanding*, *9–10*, 7–37.

Fleischman, H., & Hopstock, P. (1993). *Descriptive study of services to limited English proficient students: Volume 1: Summary of findings and conclusions*. Arlington, VA: Development Associates.

Fleming, J., & Garcia, N. (1998) Are standardized tests fair to African Americans? Predictive validity of the SAT in black and white institutions. *The Journal of Higher Education*, *69*, 472–495.

Freedle, R. (Ed.) (1979). *New directions in discourse processing*. Norwood, NJ: Ablex.

Freedle, R. O., & Duran, R. P. (Eds.) (1987). *Cognitive and linguistic analyses of test performance*. Norwood, NJ: Ablex.

Gal'perin, P. Ya. (1966). Kucheniui ob interiorizatsii [Toward the theory of interiorization]. *Voprosky Psikhologii*, *6*, 20–29.

Gándara, P., & Maxwell-Jolly, J. (1998). *Priming the pump: A review of programs that aim to increase the achievement of underrepresented minority undergraduates*. New York: College Board.

Garcia, E. (1994). *Understanding and meeting the challenge of student cultural diversity*. Boston: Houghton Mifflin.

Garcia, G. (1988). *Factors influencing English reading test performance of Spanish-English bilingual children*. Unpublished doctoral dissertation, University of Illinois, Urbana-Champaign.

Ginzburg, M. P. (1981). O vozmozhnoi interpretatsii poniatia zony blizhaishego razvitia [On a possible interpretation of the concept of the zone of proximal development]. In D. B. El'konin & A. L. Venger (Eds.), *Diagnostika uchebnoi diatel'nosti i intellectual'nogo razvitia detei* (pp. 145–155). Moscow: Academia Pedagogicheskikh Nauk SSSR.

Goncharova, E. I. (1990). Nekotorye voprosy vyshego obrazovania vzroslykh slepoglukikh [On higher education for the deaf-blind]. In V. N. Chulkov, V. I. Lubovsky, & E. N. Martsinovskaia (Eds.), *Differensirovannyi podkhod pri obuchenii I vospitanii slepoglukhikh detei* (pp. 56–70). Moscow: Academia Pedagogicheskikh Nauk SSSR.

Gonzalez, V. (1996). Do you believe in intelligence? Sociocultural dimensions of intelligence assessment in majority and minority students. *Educational Horizons*, 45–52.

Gonzalez, V., Bauerle, P., & Felix-Holt, M. (1996). Theoretical and practical implications of assessing cognitive and language development in bilingual children with qualitative methods. *Bilingual Research Journal*, 20, 93–131.

Gordon. E. (1995). Toward an equitable system of educational assessment. *Journal of Negro Education*, 64, 360–372.

Gordon, E. (1999). *Education and justice*. New York: Teachers College Press.

Gordon, E., & Bonilla-Bowman, C. (1994). Equity and social justice in educational achievement. In R. Berne & L. Picas (Eds.), *Outcome equity in education*. Thousand Oaks, CA: Corwin Press.

Gose, B. (1999, September 17). More points for 'strivers': The new affirmative action? *The Chronicle of Higher Education*, 46, A55–A56.

Grigorenko, E., & Sternberg. R. (1999). Dynamic testing. *Psychological Bulletin*, 124, 75–111.

Guthke, J. (1992). Learning tests: The concept, main research findings, problems, and trends. *Learning and Individual Differences*, 4, 137–151.

Guthke, J. (1993). Current trends in theories and testing of intelligence. In J. H. M. Hamers, K. Sijtsma, & A. J-J. M. Ruijssenaars (Eds.), *Learning potential testing* (pp. 13–20). Amsterdam: Swets & Zeitlinger.

Guthke, J., Beckman, J. F., & Dobat, H. (1997). Dynamic testing: Problems, use, trends, and evidence of validity. *Educational and Child Psychology, 14*, 17–32.

Hakuta, K., & Valdes, G. (1994). *A study design to evaluate strategies for the inclusion of LEP students in the NAEP state trial assessment*. Washington, DC: National Center for Education Statistics.

Haney, W., & Scott, L. (1987). Talking with children about tests: An exploratory study of test item ambiguity. In R. O. Freedle & R. P. Duran (Eds.), *Cognitive and linguistic analyses of test performance* (pp. 298–368). Norwood, NJ: Ablex.

Harris-Stefanakis, E. (2000). Teacher's judgments do count: Assessing bilingual students. In Z. Beykont (Ed.), *Lifting every voice: Pedagogy and politics of bilingualism* (pp. 139–160). Cambridge: Harvard Education Publishing Group.

Hessels, M. G .P., & Hamers, J. H .M. (1993). The learning potential test for ethnic minorities. In J. H. M. Hamers, K. Sijtsma, & A. J-J. M. Ruijssenaars (Eds.), *Learning potential testing* (pp. 13–20). Amsterdam: Swets & Zeitlinger.

Heubert, J., & Hauser, R. (Eds.), (1999). *High stakes: Testing for tracking, promotion, and graduation*. Washington, DC: National Academy Press.

Hill, C. (1974). Spatial perception and linguistic encoding: A case study in Hausa and English. *Studies in African Linguistics, 4*, 135–148.

Hill, C. (1975a). Variation in the use of 'front' and 'back' by bilingual speakers. *Proceedings of the Berkeley Linguistics Society, 1*, 196–206.

Hill, C. (1975b). Sex-based differences in cognitive processing of spatial relations by bilingual students in Niger. In R. Herbert (Ed.), *Patterns of language, culture, and society* (Ohio State University Working Papers, 19, pp.185–198). Columbus: Ohio State University.

Hill, C. (1977a). A review of the language deficit position: Sociolinguistic and psycholinguistic perspectives. (*IRCD Bulletin 4*). New York: Teachers College, Columbia University, Institute of Urban and Minority Education.

Hill, C. (1977b). Urban minority students, language, and reading. *ERIC/CU Urban Diversity Series, 51 & (1975).CAL/ERIC/CLL Series on Languages and Linguistics 55* (monograph published jointly by ERIC Clearinghouse on Urban Education, Teachers College, Columbia University and ERIC Clearinghouse on Languages).

Hill, C. (1978). Linguistic representation of spatial and temporal orientation. *Proceedings of the Berkeley Linguistics Society, 4*, 524–538.

Hill, C. (1982). Up/down, front/back, left/right: A contrastive study of Hausa and English. In J. Weissenborn & W. Klein (Eds.), *Here and there. Cross-linguistic studies on deixis and demonstration*. Amsterdam: John Benjamins.

Hill, C. (1991a). Recherches interlinguistiques en orientation spatiale [Interlinguistic research in spatial orientation]. *Communications*, 53, 171–207.

Hill, C. (1991b). *Spatial and temporal orientation: African and African American continuities*, (LC Report 1991-1) Teachers College, Columbia University, Literacy Center.

Hill, C. (1992). *Testing and assessment: An ecological approach* (Inaugural Lecture for the Arthur I. Gates Chair in Language and Education). New York: Teachers College, Columbia University.

Hill, C. (1995). Testing and assessment: An applied linguistics perspective. *Educational Assessment*, 2, 179–212.

Hill, C. (1996). *Exemplar essay project: Theory of knowledge course*. Cardiff, Wales: International Baccalaureate.

Hill, C. (1999a). A national reading test for fourth graders: A missing component in the policy debate. In B. Z. Presseisen (Ed.), *Teaching for intelligence I*, (pp. 129–152). Arlington Heights, IL: Skylight.

Hill, C. (1999b). Constructivist assessment in early childhood education. In A. L. Costa (Ed.), *Teaching for intelligence II* (pp. 211–230). Arlington Heights, IL: Skylight.

Hill, C., Anderson, L., Watt, Y., & Ray, S. (1989). *Reading assessment in adult education: Local detail versus textual gestalt* (LC Report 89-2). New York: Teachers College, Columbia University, Literacy Center.

Hill, C., & Larsen, E. (1983). *What reading tests call for and what children do*. (Final Report for NIE Grant G-78-0095). Washington DC: National Institute of Education.

Hill, C., & Larsen, E. (1992). *Assessment in secondary education: A critical review of emerging practices*. Berkeley, CA: National Center for Research on Vocational Education.

Hill, C., & Larsen, E. (2000). *Children and reading tests*. Stamford, CT: Ablex.

Hill, C., & Parry, K. (1988). *Reading assessment: Autonomous and pragmatic models of literacy* (LC Report 88-2). New York: Teachers College, Columbia University, Literacy Center.

Hill, C., & Parry, K. (1989). Autonomous and pragmatic models of literacy: Reading assessment in adult education. *Linguistics and Education*, 1, 233–83.

Hill, C., & Parry, K. (1992). The test at the gate: Models of literacy in reading assessment. *TESOL Quarterly*, 26, 433–461.

Hill, C., & Parry, K. (1994). *From testing to assessment: English as an international language*. Harlow, UK: Longman.

Hill, C., & Pike, L. (1985). *A comparison of cloze substitutions in English texts by native speakers of English, Spanish, and Japanese*. Unpublished manuscript.

Hispanic and Asian populations expand. (2000, August 30). *The New York Times*, p. A16.

Ho, W. (1998). *A crossnational and intergenerational study of deixis*. Unpublished manuscript.

Holmes, D., Hedlund, P., & Nickerson, B. (2000). *Accommodating English language learners in state and local assessments*. Washington, DC: National Clearinghouse for Bilingual Education.

Hopstock, P., & Bucaro, G. (1993). *A review and analysis of estimates of the LEP student population*. Arlington, VA: Development Associates.

Houser, J. (1995). *Assessing students with disabilities and limited English proficiency* (Working paper 95-13). Washington, DC: National Center for Education Statistics.

Ingulsrud, J. E. (1988). *Testing in Japan: A discourse analysis of reading comprehension items*. Unpublished doctoral dissertation, Teachers College, Columbia University, New York.

Isma'il, T. (1979). *Cross-cultural variation in spatial and temporal constructs*. Unpublished doctoral dissertation, Teachers College, Columbia University, New York.

Jencks, C., & Phillips, M. (Eds.). (1998). *The black-white test score gap*. Washington, DC: Brookings Institution Press.

Ji, J. (1998). *Spatiotemporal orientation among Chinese-speaking immigrants in metropolitan New York*. Unpublished doctoral dissertation, Teachers College, Columbia University, New York.

Johnston, P. (1984). *Reading comprehension assessment: A cognitive basis*. Newark, DE: International Reading Association.

Kim, S. (1996). *Spatiotemporal processing in Korean and English: Semantic and pragmatic perspectives*. Unpublished doctoral dissertation, Teachers College, Columbia University, New York.

Kyllonen, P., & Christal, R. (1990). Reasoning ability is (little more than) working memory capacity? *Intelligence, 14*, 394–434.

Labov, W. (1972). *Language in the inner-city: Studies in the Black English vernacular*. Philadelphia: University of Pennsylvania Press.

Lanning, N. (1999, November 15). Tales out of school. New York, pp. 32–37.

Langer, J. A. (1987). The construction of meaning and the assessment of comprehension: An analysis of reader performance on standardized test items. In R. O. Freedle & R. P. Duran (Eds.), *Cognitive and linguistic analyses of test performance* (pp. 225–244). Norwood, NJ: Ablex.

Lave, J. (1988). *Cognition in practice*. Cambridge, UK: Cambridge University Press.

Lemann, N. (1999a). *The big test: The secret history of the American meritocracy*. New York: Farrar, Straus, and Giroux.

Lemann, N. (1999b, September 6). Behind the SAT. *Newsweek*, pp. 52–57.

Lin, H. S. (1993). *A TOEFL coaching school in Taiwan*. Unpublished doctoral dissertation, Teachers College, Columbia University, New York.

Lucas, S. (1998). *Prominent explanations and potential prominent factors in the black/white test score gap*. Washington DC: Board on Testing and Assessment.

Mahmoud, H. (1997). *Spatial orientation among Sara and Maba speakers in the Republic of Chad*. Unpublished doctoral dissertation, Teachers College, Columbia University, New York.

Menken, K. (2000). *What are the critical issues in wide-scale assessment of English language learners?* (NCBE Issue Brief 6). Washington, DC: National Clearinghouse for Bilingual Education.

McKenna, S. (1985). *Cross-cultural variation in the use of locative constructs: A case study in metropolitan New York*. Unpublished doctoral dissertation, Teachers College, Columbia University, New York.

Mooney, W., & Goldstein, L. (1982). Pilot study in a Harlem high school: Language variation based on an interlocutor shift. Unpublished manuscript.

Miller, L. (1995). *An American imperative: Accelerating minority educational advancement*. New Haven, CT: Yale University Press.

NAACP seeks to limit use of College Board tests. (1999, November 21). *The New York Times*, p. A23.

Nix, D., & Schwartz, M. (1979). Toward a phenomenology of reading comprehension. In R. O. Freedle (Ed.), *New directions in discourse processing* (pp. 183–196). Norwood, NJ: Ablex.

Ogbu, J. (1994). Understanding cultural diversity and learning. *Educational Researcher*, *21*, 5–14.

Olson, J. (1995). *Puerto Rico assessment of educational progress: Final report, 1994 mathematics assessment*. Princeton, NJ: Educational Testing Service.

Olson, J., & Goldstein, A. (1996). *Focus on NAEP: Increasing the inclusion of students with disabilities and limited English proficiency* (NCES 96-894). Washington, DC: National Center for Education Statistics.

Olson, J., & Goldstein, A. (1997). *The inclusion of students with disabilities and limited English proficient students in large-scale assessments* (NCES 97-482). Washington, DC: National Center for Education Statistics.

Parry, K. (1986). *Readers in context: A study of northern Nigerian students and School Certificate texts*. Unpublished doctoral dissertation, Teachers College, Columbia University, New York.

Pearson, P. D., & Valencia, S. (1987). Assessment, accountability, and professional prerogative. In J. E. Readence & R. S. Baldwin (Eds.), *Research in literacy: Merging perspectives* (pp. 3–16). Rochester, NY: National Reading Conference.

Powers, D., & Rock, D. (1998). *Effects of coaching on SAT I: Reasoning scores* (College Board Report 98-8). New York: College Entrance Examination Board.

Presseisen, B. Z. (Ed.) (1999). *Teaching for intelligence I*. Arlington Heights, IL: Skylight.

Ramist, L., Lewis, C., & McCamley-Jenkins, L. (1994). *Student group differences in predicting college grades: Sex, language, and ethnic groups* (College Board Report 93-1). New York: College Entrance Examination Board.

Reyes, M. (1994). *Variation in the use of orientational terms among Puerto Rican students*. Unpublished doctoral dissertation, Teachers College, Columbia University, New York.

Rivera, C., & Stansfield, C. (1998). Leveling the playing field for English language learners: Increasing participation in state and local assessments through accommodations. In R. Brandt (Ed.), *Assessing student learning: New rules, new realities*. Arlington, VA: Educational Research Service.

Rivera, C., Vincent, C., Hafner, A., & LaCelle-Peterson, M. (1997). *Statewide assessment programs: Policies and practices for the inclusion of limited English proficient students—Findings from a national survey of state assessment directors*. Washington, DC: ERIC Clearinghouse on Assessment and Evaluation.

Sapir, E. (1921). *Language*. New York: Harcourt, Brace, & World.

Schemo, D. (2000, October 14). U. of Maryland branch is beacon for minorities in math and science. *The New York Times*, p. A12.

Schemo, D. (2001, February 17). Head of U. of California seeks to end SAT use in admissions. *The New York Times*, pp. A1, A13.

Secada, W. (1994). Issues in the development of Spanish-language versions of the National Assessment of Educational Progress. In G. Bohrnstedt (Ed.), *The trial state assessment: Prospects and realities*. Palo Alto, CA: Armadillo Press.

Schwartz, T. (1999, January 10). The test under stress. *New York Times Magazine*, pp. 30–63.

Sebrechts, M., Bennett, R., & Rock, D. (1991). Agreement between expert system and human raters' scores on complex constructed-response quantitative items. *Journal of Applied Psychology*, 76, 856–862.

Shepard, L. (1996). *Research framework for investigating accommodations for language minority students*. Los Angeles: UCLA Center for the Study of Evaluation/National Center for Research on Evaluation, Standards, and Student Testing.

Sims-West, N. E. (1996). *An investigation of gender differences on the Scholastic Aptitude Test of verbal ability*. Unpublished doctoral dissertation, Teachers College, Columbia University, New York.

Smitherman, G. (1975). *Talkin' and testifyin'*. Boston: Houghton-Mifflin.

Snow, R., & Lohman, D. (1989). Implications of cognitive psychology for educational measurement. In R. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 263–332), New York: Macmillan.

Stancavage, F., Godlewski, C., & Allen, J. (1996). Study of exclusion and accessibility of students with limited English proficiency in the 1994 trial state assessment of the National Assessment of Educational Progress. *In Quality and utility: The 1994 Trial State Assessment in reading background studies*. Stanford, CA: National Academy of Education.

Steele, C. (1997). A threat in the air—How stereotypes shape intellectual identity and performance. *American Psychologist*, *52*, 613–629.

Steele, C. (1999). Thin ice: "Stereotype threat" and Black college students. *Atlantic Monthly*, *284*, 44-47.

Steele, C., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*, *69*, 797–811.

Stefanakis, E. (1998). *Whose judgment counts*? Portsmouth, NH: Heinemann.

Stewart, D., & Everson, H. (1994). Educational assessment and national standards: The equity imperative. In M. Nettles (Ed). *Equity and assessment* (pp. 263–272). Boston: Kluwer Academic Press.

Valdes, G., & Figueroa, R. (1994). *Bilingualism and testing: A special case of bias*. Norwood, NJ: Ablex.

Vlasova, T. A. (1972). New advances in Soviet defectology. *Soviet Education*, *14*, 20–39.

Vygotsky, L. S. (1962). *Thought and language*. Cambridge, MA: MIT Press. (Original work published 1934)

Vygotsky, L. S. (1978). *Mind in society*. Cambridge, MA: Harvard University Press.

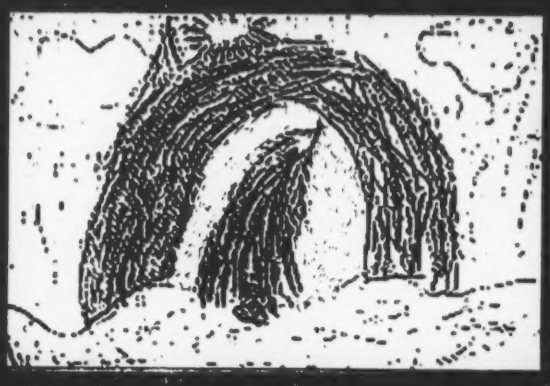
Wang, H., & Hill, C. (1999). Short-answer questions in College English testing. In Y. Teng (Ed.), *Second language acquisition and English learning strategies* (pp. 317–338). Nanjing, China: Nanjing University Press.

Wei, Y. (1994). *Linguistic representation of spatial and temporal orientation: Structured variation among Chinese university students*. Unpublished doctoral dissertation, Teachers College, Columbia University, New York.

Whorf, B. (1956). Language, thought, and reality. In J. Carroll (Ed.), *Selected writings of Benjamin L. Whorf*. Cambridge, MA: MIT Press.

Yuan, Y. P. (1997). *Reader response to the Taiwan Joint College Entrance Examination: English Reading Section*. Unpublished doctoral dissertation, Teachers College, Columbia University, New York.

Zehler, A., Hopstock, P., Fleischman, H., & Greniuk, C. (1994). *An examination of assessment of limited English proficient students*. Arlington, VA: Development Associates.



UCLA Graduate School of Education & Information Studies